Univerzita Karlova v Praze
Přírodovědecká fakulta

Studijní program: Chemie (N1407)
Studijní obor: Modelování chemických vlastností nano- a biostruktur

Bc. Dávid Jakubec

Interakční preference v komplexech protein – DNA
Interaction preferences in protein – DNA complexes

Diplomová práce

Školitel: RNDr. Jiří Vondrášek, CSc.

Praha, 2015

Prohlášení:

Prohlašuji, že jsem závěrečnou práci zpracoval samostatně a že jsem uvedl všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V Praze, 11.5.2015

Podpis

Acknowledgements

I would like to express my gratitude to my supervisor, Dr. Jiří Vondrášek, for providing countless advice and guidance thourough this work; to Dr. Roman A. Laskowski for aid with the retrieval of the data set; to Jiří Hostaš for his ongoing involvement in the calculation of benchmark results, and to the entire Bioinformatics group at the Institute of Organic Chemistry and Biochemistry, AS CR, for creating a friendly, inspirative and supportive environment.

Abstrakt

Interakce proteinů s DNA jsou základem mnoha esenciálních biologických pochodů. Navzdory dosavadním snahám se zatím nepodařilo kompletně objasnit pravidla řídící rozpoznávání specifických úseků nukleových kyselin proteiny. V této práci se pokouším prozkoumat proces rozpoznávání DNA rozdělením složité sítě kontaktů na rozhraní protein – DNA do příspěvků jednotlivých párů aminokyselina – nukleotid. Tyto páry byly získány z existujících struktur protein – DNA komplexů ve vysokém rozlišení a zpracovány bioinformatickými metodami a nástroji výpočetné chemie. Nově jsem zavedl kritéria specificity sprahující pozorované geometrické preference s relativní energetickou bilancí párů. Aplikací těchto kritérií jsem rozšířil knihovnu párů aminokyselina – nukleotid které se mohou podílet na přímém rozpoznávání sekvence. S cílem prozkoumat fyzikální základy pozorované specificity jsem vypočítal mapy elektrostatických potenciálů pro jednotlivé nukleotidy a vybrané komplexy.

Abstract

Interactions of proteins with DNA lie at the basis of many fundamental biological processes. Despite ongoing efforts, the rules governing the recognition of specific nucleic acid sequences have still not been universally elucidated. In this work, I attempt to explore the recognition process by splitting the intricate network of contacts at the protein – DNA interface into contributions of individual amino acid – nucleotide pairs. These pairs are extracted from existing high-resolution structures of protein – DNA complexes and investigated by bioinformatics and computational-chemistry based methods. Criteria of specificity based on the coupling of observed geometrical preferences and the respective interaction energies are introduced. The application of these criteria is used to expand the library of amino acid – nucleotide pairs potentially significant for direct sequence recognition. Electrostatic potential maps are calculated for individual nucleotides as well as for selected complexes to investigate the physical basis of the observed specificity.

Contents

1 Introduction **9**

2 Methods **14**
2.1 Data set preparation 14
2.1.1 Addressing data set bias 17
2.1.2 Extraction of contacts with DNA bases 21
2.2 Interaction energy calculations 22
2.2.1 Empirical methods 23
2.2.2 System partitioning and computational execution . . . 30
2.3 Electrostatic potentials 35

3 Results **37**
3.1 Evaluation of force field performance 37
3.2 Large-scale binding preferences 45
3.2.1 Interactions directed at the DNA bases 47
3.2.2 Interactions with the DNA backbone 54
3.3 Electrostatic potentials 63

4 Discussion **68**
4.1 Gas phase approximation 68
4.2 Pair-wise approximation and many-body effects 70
4.3 Comparison with protein – protein interactions 73

5 Conclusions **75**

Bibliography **77**

Appendix **89**

List of publications

Some of the results presented in this thesis have already appeared in the article *"Large-Scale Quantitative Assessment of Binding Preferences in Protein – Nucleic Acid Complexes"* published in the Journal of Chemical Theory and Computation, volume 11(4), pp. 1939–1948, on March 19, 2015.

List of abbreviations

dNMP . . . . 2'-deoxyribonucleoside 5'-monophosphate
dAMP . . . . deoxyadenosine 5'-monophosphate
dCMP . . . . deoxycytidine 5'-monophosphate
dGMP . . . . deoxyguanosine 5'-monophosphate
CBS . . . . . complete basis set
DNA . . . . . deoxyribonucleic acid
FF . . . . . force field
HF . . . . . Hartree-Fock
MAD . . . . . median absolute deviation
MD . . . . . molecular dynamics
MM . . . . . molecular mechanics
MP . . . . . Møller-Plesset (perturbation theory)
MSE . . . . . mean signed error
NMR . . . . . nuclear magnetic resonance
PDB . . . . . Protein Data Bank
PMF . . . . . potential of mean force
QM . . . . . quantum mechanics
RNA . . . . . ribonucleic acid
RMSD . . . . root mean square deviation
SCF . . . . . self-consistent field
TMP . . . . . thymidine 5'-monophosphate

Chapter 1

Introduction

Interactions of proteins with nucleic acids are essential for fundamental processes of cellular physiology. These interactions involve both the deoxyribonucleic acid (DNA) during the replication and transcription of genomic material, as well as ribonucleic acid (RNA) in post-transcriptional regulation and translation of genetic information.

Binding of proteins to DNA can display various levels of specificity towards the designated DNA sequences. In eukaryotes, the majority of untranscribed DNA is bound to histone proteins in the form of nucleosomes [1]. Although their genome-wide positionining shows sequence-dependent preferences, these particles contain regions which promote non-specific interactions with the nucleic acid [2–4]. Likewise, interactions of DNA with some repair enzymes must display low sequence specificity if genome integrity is to be maintained [5–7].

For other processes, such as the regulation of gene expression, DNA sequence recognition with high specificity is critical. Crystallographic and nuclear magnetic resonance (NMR) experiments have been actively used to explore atomic-level details of free nucleic acids and proteins, beginning with the discovery of DNA structure by Watson and Crick [8]. Repositories such as the Protein Data Bank (PDB) currently house over 3,000 structures of protein – DNA complexes obtained by a variety of experimental methods [9]. Recently, large libraries of proteins in complex with their cognate DNA motifs have been generated for some organisms [10, 11]

Sequence-specific binding of proteins to DNA can be recognised experimentally by a large decrease of the standard heat capacity of the system. This is caused by the restriction of configurational degrees of freedom of the interacting partners, as well as by features characteristic of other specific processes, such as the burial of hydrophobic residues. Thermodynamically, this binding can be driven by either enthalpic or entropic contributions, de-

pending on the temperature [12, 13]. Although most DNA-binding proteins also interact with the DNA non-specifically, this binding is not associated with the mentioned heat capacity decrease and is driven by enthalpy [13].

Structural biology and bioinformatics are two fields that have emerged in response to the growing amounts of experimental data, utilising the power of modern computational technology to discover the underlying principles of biological processes. Understanding the rules governing specific DNA sequence recognition by proteins is one of their primary goals. Despite the ongoing efforts, no recognition code applicable to interactions of all protein families has been described to date [14]. The biomedical potential of having complete control over the genomic material is enormous. The few proteins (zinc-finger nucleases and transcription activator-like effector proteins) whose DNA-binding domains can be designed to target a specific DNA sequence according to a simple amino acid – nucleotide matching code have immediately found use in genetic engineering [15].

Years of analyses of a large number of experimental structures of protein – DNA complexes have revealed two principial modes utilised in specific sequence recognition. Base readout involves local interactions between a protein DNA-binding domain and the target DNA sequence, typically in the form of a matching pattern of hydrogen bond donor and acceptor groups [14]. The possibility of amino acids recognising individual DNA bases by bidentate hydrogen bond contacts was first explored by Seeman based on early structural data [16]. It was realised that asparagine and glutamine are capable of uniquely distinguishing between adenine and the other bases in the major groove, while a specific recognition of guanine by these amino acids is possible in the minor groove. In addition, arginine can be used to recognised guanine in the major groove [16].

Moving onto the protein secondary structure level, the most common motif *via* which the protein interacts with the DNA is an α-helix inserted into the major groove [14]. This helix can be a part of a larger supersecondary structure, for example, a helix-turn-helix motif utilised by the ETS domains [17]. Interactions of DNA with the β-sheet structures have also been observed, although they often result in a significant deformation of the DNA molecule [18].

It was soon realised that this linear picture of specific DNA sequence recognition describing only local features of the interaction interface was not complete. The readout of the DNA shape was found to be equally important in some complexes [14]. Non-canonical forms of the nucleic acid have been described in many protein – DNA structures [18–20]. The predisposition to form various local deviations is known to be dependent on the DNA sequence and varies between different regions of the genome [21–23]. For example, GC-

rich sequences have higher propensity to assume A-like forms [24]. On the other hand, a narrowing of the minor groove often observed in AT-rich regions creates more negative electrostatic potential, which is universally recognised by arginine side chains [23]. A global bend of the DNA structure induced by the interaction with proteins can enable the formation of contacts that would be impossible with free DNA [18, 22, 25].

Base readout and DNA shape recognition are two extremes that are usually combined in real protein – DNA complexes. The binding of proteins can induce a conformational change in the nucleic acid, which may, in turn, enable the formation of a new sets of contacts. Therefore, the two interaction modes are not independent and can not be separated if a complete description of the recognition proces is to be provided. Based on an analysis of a large amount of structures of protein – DNA complexes, it was concluded that the while the motifs involved in base readout can distinguish between individual families of DNA-binding proteins, the niche differences in the dynamic properties of the cognate DNA region can guide the higher-resolution recognition by specific members of a single protein family [14, 22, 26].

The DNA shape recognition depends on non-local dynamic properties of larger DNA residue blocks which are difficult to generalise across different sequences [23]. On the other hand, studies of amino acid – DNA base pairs which probe the direct base readout mechanism have been readily performed. The advent of computer technology has enabled analyses which investigate the binding mechanism in thousands of protein – DNA structures at the same time. Indeed, while only limited experimental data on the interactions of amino acid – DNA base pairs are available [27, 28], substantial part of the studies performed on these dimers has utilised bioinformatics and other computational approaches [29–31].

Mandel-Gutfreund and Margalit were among the first to utilise a library of three-dimensional structures to derive contact potentials for the prediction of protein – DNA interactions [29]. These potentials were derived by comparing the observed number of respective amino acid – DNA base pairs to that expected for a theoretical distribution and calculating the logarithm of the odds. It was found that pairs which carried complementary patterns of hydrogen bond donor and acceptor group and, therefore, enabled recognition of single DNA bases by single amino acids, were strongly favourised at the interface [29].

Luscombe *et al.* investigated the atomic-level details of the interaction interfaces of 129 structures of protein – DNA complexes. They observed significant correlations between the populations of individual DNA bases and amino acid side chains which enable their specific recognition in a one-to-one fashion. The populations of various binding motifs (van der Waals contacts,

hydrogen bonds and water-mediated interactions) involving different parts of the nucleotide were compared. It was found almost two thirds of contacts featured in direct base readout involved bidentate hydrogen bonds. However, two thirds of all interactions were found to be realised by van der Waals contacts, often with the DNA backbone, suggesting their non-specific character. Again, pairs which enabled significant one-to-one recognition of DNA bases by amino acids were favoured. In addition, some other pairs which did not involve bidentate hydrogen bonds were dubbed "context-dependent", as they were not able to uniquely distinguish between individual bases, but were clearly essential in the stabilisation of the respective complexes in which they were found [30].

Multiple online database have been established which focus on different aspects of protein – DNA interactions. The above described work by Luscombe *et al* was accompanied by a web server which contains the structures of amino acid side chain – DNA base pairs extracted from high-resolution structures of protein – DNA complexes [30]. The "Amino Acid – Nucleotide Interaction Database" was established by Hoffman *et al* and provides very similar information, but also includes contacts featuring the protein backbone residues. Both of these databases show clustering of amino acid residues in certains regions around the DNA nucleotides (see below). The "Protein – DNA Interface database" and "3D-footprint" databases offer various search criteria and analytical tools, such as browsing by protein families or visualisation of the network of contacts at the interaction interface [32, 33]. Thermodynamic data on protein – DNA complexes are summarised in the ProNIT database [34]. CollecTF and TRANSFAC are databases containing information about transcription factors found in prokaryotic and eukaryotic organisms, respectively [35, 36]. Finally, general databases such as the "Nucleic acid – Protein Interaction DataBase" provide some enhanced functionality and tools focused on protein – DNA complexes compared to the PDB [9, 37].

The studies of base readout presented so far have all focused on statistical analysis of existing three dimensional structures. These analyses, however, do not explicitly investigate the physico-chemical characteristics of the interacting partners. A different approach, based on the methods of computational chemistry, is possible. Indeed, theoretical studies calculating the properties of amino acid – DNA base pairs have been conducted by both quantum mechanical (QM) as well as empirical methods.

Molecular electrostatic potentials of isolated DNA bases and DNA base pairs were calculated from self-consistent field (SCF)-level wave functions already in early 1970s. [38, 39]. Šponer and Hobza demonstrated with *ab initio* calculations that amino groups of DNA bases adapt non-planar geometries when electron correlation energy is included [40]. Hobza and Šponer also

calculated the accurate stacking energies of various DNA base dimers from first principles by extrapolating the results of the coupled clusters calculations covering single and double excitations iteratively and triple excitations perturbatively (CCSD(T)) to the complete basis set (CBS) limit [41]. Accurate energies of hydrogen-bonded nucleic acid base pairs were determined in a similar way and deposited in a benchmark database [42, 43]. Accurate CCSD(T)/CBS interaction energies of amino acid – DNA base pairs have been calculated by Hostaš *et al.* (manuscript submited), while calculations on amino acid – DNA nucleotide dimers are currently being performed.

Very recently, absolute binding free energies of amino acid – DNA base pairs in aqueous and methanol environments were calculated by de Ruiter and Zagrovic [44]. These were determined by calculating the potentials of mean force (PMF) obtained from molecular dynamics (MD) simulations. Free energy maps of amino acid – DNA base interactions were calculated by Pichierri *et al.* from the partition function [45].

In this work, the approach based on statistical analyses of three-dimensional structures of will be combined with empirical calculations. Contacts of amino acid side chains with the DNA bases will be extracted from a large set of high-quality structures of protein – DNA complexes. The energetical contribution of various pair geometries to the base readout mechanism will be investigated. Afterwards, the sugar-phosphate moieties will be added to the DNA bases to study the effects of the negatively charged group on the interaction specificity. The energetics of contacts with the DNA backbone atoms will then be considered and compared with the arrangements involved in base readout. Criteria coupling geometrical preferences of the amino acid side chain – DNA residue pairs to the large-scale energetic characteristics of the respective amino acid – DNA base pair combinations will be defined in an attempt to discover specific binding motifs which could not be seen in the previous studies. This work will also use the unique opportunity to test the reliability of three commonly used molecular mechanical (MM) force fields (FF) by comparing the respective interaction energies with the results of accurate CCSD(T)/CBS calculations. Finally, electrostatic potentials will be calculated for multiple molecules and molecular complexes to examine the physical basis of the interactions.

Chapter 2

Methods

This section begins with the description of construction of the data set, that is, the extraction of amino acid – DNA nucleotide pairs from the available structures of protein – DNA complexes. Geometrical similarity of some pairs is noted and rigorously defined. As many of the protein structures are homologous and their inclusion would introduce bias, Section 2.1.1 addresses the treatment of the redundant entries. In Section 2.1.2, a subset of structures in which the amino acid interacts with the DNA base moiety is constructed from the larger set of all contacts. The methodology of interaction energy calculations is introduced in Section 2.2, beginning with a brief overview of the benchmark *ab initio* method. Afterwards, a detailed description of MM force fields is provided, with focus on the derivation of parameters which determine molecular properties in non-covalent complexes (Section 2.2.1). Nuances of the particular computational execution are described in Section 2.2.2. Finally, in Section 2.3, the theory behind the calculation of electrostatic potentials is explained.

2.1 Data set preparation

The structural data used as a basis of my work were obtained in collaboration with the author of the "Atlas of Protein Side-Chain Interactions". The atlas is accessible online at `http://www.ebi.ac.uk/thornton-srv/-databases/sidechains/` and is based on a printed (1992) book of the same name [46]. This printed version contains analyses of the geometries of amino acid pairs extracted from 62 crystal structures of proteins obtained from the Brookhaven (now RSCB) Protein Data Bank [9]. It provides illustrations of various interaction motifs supplemented with their relative populations, as well as the histograms summarising the mutual orientations of the interacting

partners in spherical coordinates [46, 47].

The online version of the printed atlas was established in 2001, utilising the by then greater amount of available protein structures to extend the library used for the atlas construction[1]. In addition, the web version included a section containing analyses of contacts found in protein – DNA complexes. These contacts — amino acid – 2'-deoxyribonucleoside 5'-monophosphate (dNMP) dimers — were originally extracted from the structures of 192 complexes of proteins with DNA solved to a resolution higher than 3.0 Å in which the total length of the double stranded DNA region was at least 4 base pairs [30]. These complexes were also obtained from the Protein Data Bank [9].

The web version of the atlas utilises the SIRIUS set of Fortran scripts [47] to extract the amino acid – dNMP dimers from the corresponding protein – DNA complexes. First, a set of atoms which serve as points of reference is defined for each of the 20 standard amino acids. These are usually side chain functional group atoms that are characteristic for each amino acid (*i.e.*, cysteine thiol group sulfur atom). Main chain C_α atom is used for glycine. An amino acid – dNMP is recognised as interacting when the distance between any of the amino acid reference atoms and any dNMP heavy (non-hydrogen) atom is less than the sum of their van der Waals' radii plus 1.0 Å [47].

As the atlas web server had not been actively maintained since 2006, we had to perform the extraction of the interacting amino acid – dNMP dimers from the up-to-date list of available structures manually. As of March 2014, there had been a total of 3,143 structures of protein – DNA complexes deposited in the PDB [9]. From these, we obtained a high quality subset of structures solved by X-ray crystallography to a resolution better than 2.5 Å and having an R-factor[2] no worse than 0.25 using the PISCES sequence culling server [50]. This server can be used to obtain subsets of sequences from larger lists of structures based on user-defined criteria, including percentage sequence identity, resolution, chain length and experimental method [50]. Only structures containing at least one double stranded DNA region consisting of at least 4 base pairs were considered. There were a total of 1,584 X-ray PDB entries satisfying out criteria of resolution and R-factor requirements. A single chain was considered when multiple identical polypeptide chains were included in the PDB structure, for example one of a homodimeric protein. For heteromultimeric proteins, the polypeptide chains were separated and further analysed independently (*i.e.*, during sequence homology assessment).

[1]The number of crystal structures had in the decade since the book version had been published grown by an order of magnitude [48].

[2]R-factor is a measure of agreement between the observed pattern of reflections and those calculated from the model [49].

A total of 1,737 unique polypeptide chains in complex with DNA established the data set[3].

From each of the accepted PDB structures, all amino acid – DNA nucleotide dimers were extracted by applying the distance criteria defined by the SIRIUS [47] scripts: when the distance between any amino acid reference atom and any DNA residue heavy atom was less than the sum of their van der Waals' radii (Table 2.1), a contact was recognised. The pairs in which

Element	Radius (Å)
Carbon	1.70
Nitrogen	1.55
Oxygen	1.52
Phosphorus	1.80
Sulfur	1.80

Table 2.1: The van der Waals' atomic radii values used. From [51].

the nucleotide moiety originated from the 5' end of the DNA strand were excluded, as these residues naturally lack the phosphate group. Contacts involving both *syn-* and *anti-* conformations of the 2'-deoxyribonucleoside were considered. A total of 47,480 dimers were obtained this way.

When one transforms all dimers containing a certain amino acid – dNMP pair (for example, all deoxyadenosine 5'-monophosphate (dAMP) – asparagine contacts) into a common frame of reference, a three-dimensional distribution of amino acid residues around the DNA base results (Figure 2.1). This transformation was performed by minimising the root mean square deviation (RMSD) of the nitrogenous base heavy atoms between all pairs of a particular type. The resulting distributions reflect the accessability of the DNA nucleotide base, sugar and phosphate moieties in the DNA double helix [30].

The directional nature of some interaction modes, notably hydrogen bonds, leads to the clustering of amino acid residues relative to the base in three dimensions (Figure 2.2) [30, 53]. These clusters were rigorously identified as follows. After all dimers of a certain type had been transformed to superpose the DNA bases as described above, we picked out each amino acid in turn and

[3]There are currently (April 2015) 3,316 structures of protein – DNA complexes in the PDB, of which the PISCES [50] web server returns 1,721 X-ray structures with resolution better than 2.5 Å and R-factor no worse than 0.25, comprising 1,888 unique polypeptide chains after the correction for homomultimeric proteins is applied. For comparison, when the atlas web server was last updated (October 2006), only 1,256 structures of protein – DNA complexes were available.

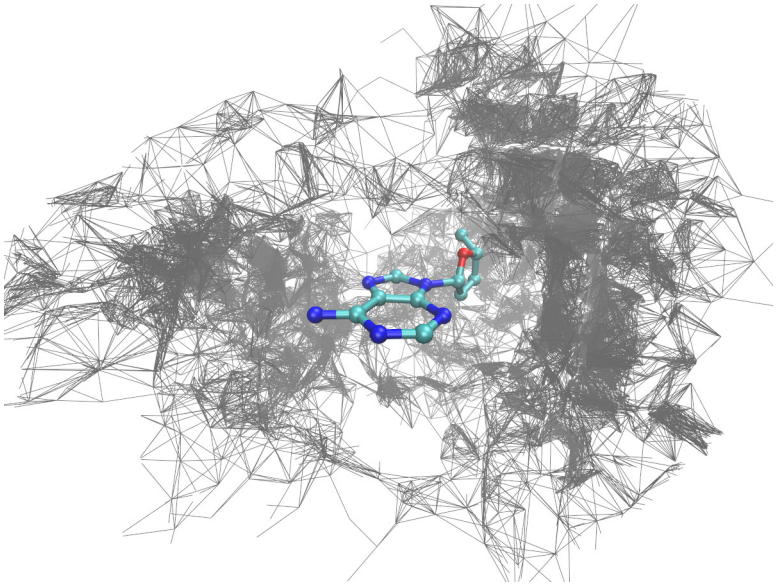


Figure 2.1: Asparagine side chain distribution (grey) around 2'-deoxyadenosine (ball-and-stick). Visualised using VMD-1.9.2 [52].

calculated the RMSD between its reference atoms and the reference atoms of all the other amino acids in the corresponding distribution. The amino acid for which the number of contacts with RMSD less than 1.5 Å was the largest was then recognised as a cluster representative and was together with its neighbours (the cluster) taken out of the distribution. The process was repeated until 6 clusters were found for each distribution, or until the last cluster isolated was too sparsely populated to be considered significant. The significance of each cluster was evaluated by assessing the probability that the cluster would emerge by chance after random rearanging of the amino acids in the distribution; if the identified cluster was smaller than an average randomly created one, it was discarded [30, 53]. A total of 12,935 dimers were found within one of the 469 clusters. Cysteine is the only amino acid for which there were found some insignificant clusters. Table 2.2 summarises the numbers of structures in clusters and distributions for individual base types.

2.1.1 Addressing data set bias

While the redundant polypeptide chains corresponding to identical protein units within individual PDB files had already been discarded, no sequence identity was investigated for entries originating from different PDB structures as of this point. This redundancy would introduce bias into the data set, as

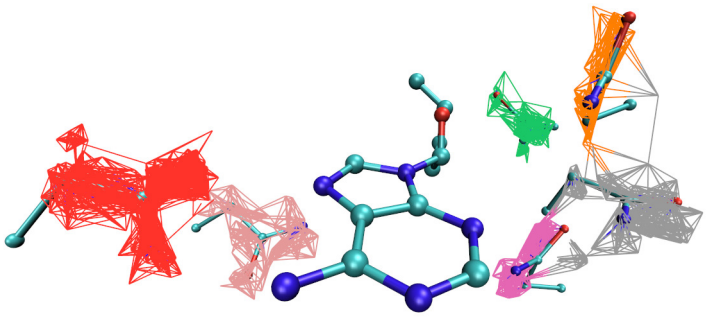


Figure 2.2: Asparagine side chains found in clusters (colourful lines) in the adenosine – asparagine distribution above. Ball-and-stick structures within clusters in correspond to cluster representatives. Visualised using VMD-1.9.2 [52].

the contacts originating from homologous protein structures would appear overpopulated compared to the contacts extracted from protein families for which few structures are currently available.

This bias was treated by first performing a global sequence alignment using the Needleman-Wunsch algorithm [54] for all pairs of protein chain sequences. The Needleman-Wunsch algorithm is a dynamic programming procedure which simplifies the complex problem of calculating the optimal global alignment of two protein or nucleic acid sequences (a function of all residues in both sequences) into a set of calculations involving only pairs of residues. A scoring matrix is used to assign a score to every residue substitution or match that can occur during the alignment; gap opening and extension penalties must be added. The best alignment is then found iteratively by following the path of the highest scores for each position in the calculated score matrix, starting from the end position [54]. BLOSUM62 substitution matrix was used, which is a *de facto* standard when comparing sequence with unknown level of similarity [55]. This matrix was constructed by calculating the logarithm of the odds of the ratio of observed to expected substitution frequencies on a set of aligned protein sequences with at most 62% mutual sequence identity [56]. The sequence alignments were carried

Number of contacts	in clusters	in distributions
Adenine	2,800	11,080
Guanine	4,037	14,467
Cytosine	3,268	10,603
Thymine	2,830	11,330
Total	12,935	47,480

Table 2.2: Number of contacts in clusters compared to the whole distribution populations for each of the DNA base types.

out using the *needle* and *stretcher* packages available in the EMBOSS-6.4.0.0 molecular biology suite [57]. Gap opening and extension penalties were the default values 10.0 and 0.5 for *needle* and 12.0 and 2.0 for *stretcher*, respectively. The *stretcher* program uses a faster variant of the Needleman-Wunsch algorithm which requires less memory and is better suited for aligning longer sequences [58]; *needle* utilises the unmodified procedure. These two tools were compared (see Figure A.1). The percentage of alignments with sequence identity at most X% is shown as a function of sequence identity assigned by the respective tools. The result of this comparison indicates that the *needle* tool, likely due to lower penalties for gap opening and elongation, finds better alignments for highly divergent sequences; however, starting at around 20% sequence identity, essentially identical alignments are obtained from both tools. Therefore, only one set of sequence identity scores was needed; the one provided by the *needle* tool was chosen.

The sequence homology was investigated at 30%, 90%, 95% and 100% sequence identity levels (X = 30, 90, 95, 100; 100% = removal of identical chains). It was expected that most of the redundancy is a result of the overpopulation of few protein families, the entries from which would already be filtered out in the 90–100% sequence identity range. The 30% sequence identity level is to be considered extreme in this manner; the corresponding set is to be considered to contain completely non-homologous proteins.

For each of the 1,737 protein chains, a list of proteins having sequence identity greater than X% was compiled. These lists were then merged for each X to create a total list of homologous structures at that particular sequence identity level; the complements of these lists are sets of structures for which the sequence identity of any pair is less than X%. These complementary sets of structures were then considered non-redundant at X% sequence identity. Neither the number of contacts each structure provides nor its resolution were considered when choosing which sequence from each set would be discarded – doing so (maximising the number of contacts obtained) would

lead to a drastic increase in the computational demands of the algorithm, as all possible combinations of sequences would have to be considered for each chain.

The procedure described leads to the removal of some entries which might be unjustified: if sequences A and B have 90% sequence identity and sequences B and C have also 90% sequence identity, sequence B is rightfully removed because of its similarity to A, while C is removed because of its similarity to a protein sequence that is no longer in the data set (the sequence identity of A and C may be less than 90%). Table 2.3 shows the number of protein chains remaining in the data set for the sequence identity levels considered after the structures removed in a manner similar to sequence C ("hard") in the example above were discarded to when they were explicitly included in the data set ("soft" approach); Figure A.2 shows this comparison for a spectrum of sequence identities. It can be seen that the differences are marginal. In other words, it appears that the blocks of homologous structures in the data set form sets of sequences which are referenced together by multiple other sequences as identical at levels X% and higher. Therefore, the event when a sequence would be excluded on the basis of a single identified identity occurs very rarely. The loss of the few sequences discarded despite

Sequence identity†	30%	90%	95%	100%
Hard	391	550	593	894
Soft‡	399	550	596	894

Table 2.3: Number of protein chains left in the data set after the various redundancy reduction criteria. † – indicates that the mutual identity of any pair of sequences in the set is less than X%. ‡ – sequences that would be removed due to sequence identity with proteins that had already been discarded ("hard" approach) were explicitly included. The total number of protein chains before any redundancy issues were addressed was 1,737.

their homologous chains having already been removed therefore appears acceptable.

Table 2.4 presents the number of amino acid – dNMP dimers left in the data set after the various bias reduction criteria had been applied. Comparing with Table 2.2 (in which no dimers were removed due to bias), one can see that more than a half (54.3%) of all contacts are removed by simply discarding the sequences that are 100% identical[4]; the number gets quickly larger

[4]As the number of protein chains discarded at this identity level (843) constitutes a similar fraction of the total number of chains (48.5%), the simplification that the number of contacts a structure provides does not influence which chain remains in the non-redundant

Sequence	30%		90%		95%		100%	
identity†	Clust.	Dist.	Clust.	Dist.	Clust.	Dist.	Clust.	Dist.
Adenine	169	2,137	355	3,087	389	3,282	964	5,200
Guanine	171	2,477	280	3,411	352	3,696	1,043	6,237
Cytosine	161	2,007	277	2,783	311	2,942	948	4,899
Thymine	208	2,305	371	3,224	407	3,398	942	5,373
Total	709	8,926	1,283	12,505	1,459	13,318	3,897	21,709

Table 2.4: Number of contacts in clusters and in the distributions for each of the DNA base types after redundant contacts had been discarded. † – indicates that the mutual identity of any pair of sequences in the set is less than X%.

for 95%, 90% and 30% sequence identities (72.0%, 73.7% and 81.2% contacts removed, respectively). The cluster populations were hit much harder by the removal of redundant chains: 70.0%, 88.7%, 90.1% and 94.5% contacts in clusters were discarded after applying 100%, 95%, 90% and 30% sequence identity criteria, respetively. This behaviour was expected, as homologous structures were more likely to contain similar geometries of amino acid – dNMP pairs. The fact that the number of sequences removed differs relatively little between 90% and 30% sequence identity levels confirms that the bias in the set is caused by several overpopulated protein families; the protein chains that originate from these entries are already discarded at the 90% sequence identity level.

2.1.2 Extraction of contacts with DNA bases

The contacts retrieval procedure described above obtained dimers in which the amino acid may be found in proximity to any of the dNMP moieties (base, 2'-deoxyribose, phosphate). The next step was to extract only the subset of contacts in which the amino acid interacts with the DNA base. To this end, I calculated the distances between all (not only reference) heavy atoms of the amino acid and all heavy atoms of the DNA base for every contact in each distribution. When any of the interatomic distances were smaller than the sum of the van der Waals' radii of the atoms plus 1.0 Å, the dimer was labelled as containing an amino acid – DNA base interaction. Values for the van der Waals' radii from Table 2.1 were used. The number of amino acid – DNA base dimers in clusters and in the distributions for each DNA base type is summarised in Table 2.5. By comparing Tables 2.5 and

set seems somewhat reasonable.

Sequence	30%		90%		95%		100%	
identity†	Clust.	Dist.	Clust.	Dist.	Clust.	Dist.	Clust.	Dist.
Adenine	125	1,080	264	1,548	281	1,643	546	2,462
Guanine	134	1,313	202	1,761	233	1,894	518	3,011
Cytosine	95	1,000	158	1,358	172	1,451	419	2,213
Thymine	146	1,359	256	1,879	277	1,980	496	2,886
Total	500	4,752	880	6,546	963	6,968	1,979	10,572

Table 2.5: Number of dimers present in clusters and in the distributions for each of the DNA base types after the redundant contacts had been discarded. Only the pairs in which the amino acid interacts with the DNA base are present. † – indicates that the mutual identity of any pair of sequences in the set is less than X%.

2.4, one can see that the discarding of the dimers in which the amino acid did not interact directly with the DNA base had a more pronounced effect on the population of the whole distributions than it did on the contacts found in clusters. The percentages of contacts retained in the clusters compared to those in the distributions are 70.5% *versus* 53.2%, 68.6% *versus* 52.3%, 66.0% *versus* 50.83% and 50.1% *versus* 48.7% for 30%, 90%, 95% and 100% sequence identity criteria, respectively. As the differences between these two numbers become the greater the more restrictive the applied bias reduction criteria are, one can speculate that the contacts found in the clusters truly represent significant interaction modes shared by different protein families.

2.2 Interaction energy calculations

The presented sets of contacts (cluster representatives, clusters, distributions) form a hierarchy of structures in which each subsequent set contains an order of magnitude more amino acid – dNMP pairs than the previous one. The method used to perform the interaction energy calculations on these sets must be reasonably accurate if a correct picture about the energetics of various interaction modes is to be obtained. On the other hand, it must at the same time be capable of processing tens of thousands of complexes, each of up to almost 60 atoms.

Ideally, one would treat the system using electronic structure methods, in which the energy and other properties are derived from the molecular wave function. Unfortunately, it is currently computationally off limits to perform the highly demanding QM calculations on this amount of similarly sized complexes in any reasonable time [59]. Rather than make a compromise

(*i.e.*, use semi-empirical methods), I decided to perform computationally less demanding MM interaction energy calculations on all contacts in all distributions. The results of these calculations were compared with the results of benchmark QM calculations performed on the limited set of cluster representatives. This set contains examples of all non-covalent interaction modes found in biomolecular complexes: hydrogen bonds, van der Waals contacts, electrostatic and dispersion interactions. Therefore, comparison with the results of the benchmark calculations allows the applicability of empirical methods to all presented complexes to be assessed [53].

The CCSD(T)/CBS method has been proven to be the most accurate method for assessing the interaction energies and geometries of non-covalent complexes of up to few tens of atoms [43, 60, 61]. It has already been used to investigate the interaction energies in pairs of amino acids, in DNA base dimers, as well as in few small non-biomolecular model complexes [41,43,62].

The accurate CCSD(T)/CBS results were obtained from Hostaš *et al.* (manuscript submitted). The calculations were performed as follows. First, a Hartree-Fock (HF) energy was calculated using a large basis set (aug-cc-pVQZ). The HF energy converges quickly to the complete basis set limit and no extrapolation is therefore needed. Correlation energy at the CBS limit was obtained from second-order Møller-Plesset perturbation theory (MP2)-level calculations performed in aug-cc-pVTZ and aug-cc-pVQZ basis sets by extrapolation according to the Halkier-Helgaker schema [63,64]. Adding this correlation term to the HF/aug-cc-pVQZ result yields the MP2/CBS energy. While both MP2 and CCSD(T) calculations are slow to converge to the CBS limit, the difference between the two energies is already converged in a smaller basis set. The final, correction, term (the difference between the MP2/CBS and CCSD(T)/CBS energies) was thus calculated using aug-cc-pVTZ basis set and added to the MP2/CBS result to yield the CCSD(T)/CBS energy [41, 43, 62]. All energies were corrected for the basis set superposition error (Jiří Hostaš, personal communication).

In the following section, the MM methods used are examined in detail. The focus on the way the empirical parameters were derived is crucial for the correct understanding and interpretation of the results obtained, especially when one deals with systems in an environment the methods were not primarily parametrised to. Then, the implementation of the missing parameters and computational details specific for the problem being solved are described.

2.2.1 Empirical methods

MM-based approaches (energy minimisation, molecular dynamics) have been well established in the realm of calculations performed on biomolecules, as

they are about the only methods capable of processing within reasonable time frames the systems which are of interest to computational biologists. These include membranes, proteins and nucleic acids both in isolated form as well as in non-covalent complexes with other macromolecules [65–69]. The system is treated using classical (Newtonian) mechanics, with the forces acting on atoms given by the gradient of an analytical potential energy function generally in the form of Equation 2.1. The potential energy is calculated by summing up terms with clear physical interpretations corresponding to intra- and intermolecular interatomic interactions (Equations 2.2–2.5):

$$E_{potential} = E_{bonds} + E_{angles} + E_{dihedrals} + E_{non-bonded} \tag{2.1}$$

where

$$E_{bonds} = \sum_{bonds} K_b(b - b_0)^2 \tag{2.2}$$

$$E_{angles} = \sum_{angles} K_\theta(\theta - \theta_0)^2 \tag{2.3}$$

$$E_{dihedrals} = \sum_{dihedrals} \sum_{n=0}^{N} \frac{V_n}{2} \left[1 + cos(n\phi - \gamma)\right] \tag{2.4}$$

and

$$E_{non-bonded} = \sum_{i} \sum_{j>i} \left\{ \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} + 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \right\} \tag{2.5}$$

This minimal funcional form is utilised by all Class I force fields (FFs), which lack terms coupling the contributions described by Equations 2.2–2.4 [70], and therefore applies to all FFs used in this study. Force field is a collective name for the functional form and its corresponding set of parameters, such as those used in Equations 2.2–2.5. MM methods are therefore empirical and their reliability is completely determined by the applicability of the parameters used to the to problem being solved [71–73].

Three combinations of FFs were used in this study: Amber99SB-ILDN protein FF with Amber99 nucleic acid parameters; Amber03 protein FF with Amber99 nucleic acid parameters and CHARMM22 protein FF with CHARMM27 nucleic acid parameters [73–78]. These FF pairs are referenced as Amber03, Amber99SB-ILDN and CHARMM27 in the text. The detailed description of these parameter sets is provided in following paragraphs.

Crucial to FF developement and modification is the concept of atom types. Atom type is an attribute assigned to atoms originating from chemically similar groups and local environments. For example, peptide bond

(amide) nitrogen and amino group nitrogen atoms have different atom types assigned in the Amber94 FF; on the other hand, all carbon atoms in aliphatic amino acids have the same atom type [73]. There is an incentive to keep the number of atom types to a minimum; the opposite would lead to over-parametrisation of the FF and limit its range of applicability. Atom types simplify the transferrability of the parameters from the molecules on which they were derived to larger systems [77]. The transferrability of parameters was important in this study when new atoms were being added to the existing molecules [53].

In Equations 2.2 and 2.3, harmonic potentials are used to describe the energy penalties for deviations from reference bond lengths and angles (parameters b_0 and θ_0, respectively). The reference bond lengths and angles are the values these terms adapt when all other energy contributions in the FF are set to zero. The force constants K_b and K_θ describe the steepness of the potential function around the minimum. In the case of Equation 2.2, one must remember that the quadratic form is only an approximation to the true bond stretching potential, which is more accurately described by an exponential function (Morse potential):

$$v(b) = D_e \left(1 - e^{-a(b-b_0)}\right)^2 \tag{2.6}$$

where

$$a = \sqrt{\frac{K_b}{2D_b}} \tag{2.7}$$

where D_b represents the depth of the bond potential energy well at minimum and K_b is the force constant of the bond at the reference bond length. The harmonic approximation is only valid when the bond length is near this reference value, which is generally true for molecules in the electronic ground state at room temperature, such as those considered in this study [71].

In Amber94 protein and nucleic acid FF, the parameters b_0, θ_0, K_b and K_θ were derived from fitting to structural data (for example, derived from X-ray diffraction) and vibrational analyses of molecular fragments. These parameters are then adjusted to reproduce the experimental normal mode frequencies [73].

Equation 2.4, the so-called torsional term, describes the energy penalties that arise as a bond is rotated. The summation runs through all sets of four atoms $\{A,\ B,\ C,\ D\}$ connected by three covalent bonds in the order A-B-C-D[5] within a single molecule. The dihedral angle is defined as the angle

[5]In addition, improper torsional angles can be defined between any quartets of atoms within a molecule, regardless of connectivity. They are often used to impose certain geometries to molecules (*i.e.*, to keep a nucleic acid base planar), notably in the CHARMM22 and CHARMM27 FFs [71, 77, 78].

between the planes containing atoms A, B, C and B, C, D. The parameters n, V_n, ω and γ stand for the bond multiplicity, potential energy barrier of the rotation, the value of the torsional angle (which is, of course, the argument of the function) and the phase factor, respectively. The bond multiplicity determines how many energy minima are encountered as the bond is rotated by 2π around the axis formed by atoms B and C. Although the general form of the torsional potential (Equation 2.4) is written as a cosine series expansion over n, most FFs use only one or few terms. The phase factor describes when does the torsional angle pass through the energy minimum [71].

In Amber94 FF, the energy associated with bond rotation is in most cases determined only by the parameters defined for the pair of atoms B and C forming the central bond. The use of only two atoms instead of four simplifies the parametrisation process and improves parameter transferability. For example, a single term in the cosine expansion with $n = 3$ was used for dihedral angles in which both B and C were sp^3 carbon atoms, except when a both A and D were highly electronegative atoms, such as oxygen or fluorine; in these cases, an additional term with $n = 2$ was included. The V_2 and V_3 parameters were determined by fitting to MP2/6-31G* energies of various conformations of simple model molecules. Additional terms were also included in the cosine expansion of the ϕ and ψ dihedral angles of the protein backbone, and χ torsional angle[6] in DNA nucleotides [73].

Several dihedral parameters were modified in the Amber98 and Amber99 FFs [74,75]. The description of the protein backbone torsional potentials was improved by deriving the dihedral parameters from *ab initio* conformational studies of alanine tetrapeptide, as compared to the alanine dipeptide used in the parametrisation of Amber94 [73,75]; these parameters were subsequently modified in the Amber99SB FF to better reproduce the relative stabilities of various protein secondary structures [79]. Amber99SB-ILDN, which was often used thorough this study, improves the description of rotameric states of side chains for isoleucine, leucine, aspartate and asparagine [80]. The parameters describing the χ torsional angle in DNA nucleotides were also modified in the Amber98 and Amber99 FFs, leading to the rotational barrier around the N-glycosidic bond being comparable with MP2/6-31G* *ab initio* and experimental results [74, 75].

In the Amber03 protein FF, the torsional parameters of the protein backbone were derived from fitting to the energies of various conformations of alanine or glycine dipeptides at the MP2/cc-pVTZ level in the presence of implicit solvent ($\epsilon = 4$). This resulted in consistency with the way the partial

[6]The torsional angle describing the rotation around the N-glycosidic bond, defined by atoms O4', C1', N9 and C2 (pyrimidines) or C4 (purines).

charges were derived (see below) [76].

CHARMM22 protein and nucleic acid FF introduces an additional term (Equation 2.8) to the potential energy function. This term is a fuction of the distance between atoms A and C (1-3 atoms) in the angle A-B-C:

$$E_{Urey-Bradley} = \sum_{angles} K_{UB}(S - S_0)^2 \tag{2.8}$$

where K_{UB} is a force constant, S is the distance between the 1-3 atoms, S_0 is the reference distance and the summations runs through all angles within a molecule. The bond stretching, angle bending and torsional term reference values (b_0, θ_0, γ and n) were optimized by fitting to microwave and electron diffraction data on gas phase structures or X-ray diffraction data on crystal structures. Only a single term from the dihedral term cosine expansion was used unless special attention was paid to that particular torsion (*i.e.*, protein backbone ϕ and ψ angles). The associated force constants and energy barries (K_b, K_θ and V_n) were fitted to reproduce the gas phase infrared and Raman vibrational spectra supplemented with 6-31G(d) level *ab initio* calculations. Finally, the Urey-Bradley parameters K_{UB} and S_0, together with improper torsional parameters, were added where the agreement with experimental data was deemed unsatisfactory [77].

In the CHARMM27 nucleic acid FF, used thorough this study, the torsional parameters of DNA nucleotides (notably the 2'-deoxyribose and phosphate moieties) were derived from fitting to MP2/6-31(+)G(d) potential energy surfaces, and subsequently refined in molecular simulations to match the experimentally observed dihedral angle distributions [78, 81].

The bond stretching, angle bending and torsional terms described so far are restricted in effect to interactions within molecules. On the other hand, the non-bonded term (Equation 2.5) applies to both intra- and intermolecular interatomic interactions. Although the summation is written over all pairs of non-identical atoms, interactions between atoms within a single molecule separated by less than three covalent bonds are usually excluded; often a scaling factor is applied to interactions between atoms separated by exactly three bonds (1-4 interactions, see below) [71].

The non-bonded term (Equation 2.5) contains contributions from electrostatic and van der Waals interactions. The electric moments (charges, dipoles, quadrupoles, ...) in the molecule are usually approximated using point partial atomic charges (parameters q_i and q_j) situated at atomic nuclei. These charges are fixed in the considered FFs, *i.e.*, no polarisation effects are explicitly included [82]. Coulomb's law is used to calculate the potential energy resulting from the electrostatic interaction between a pair of

partial atomic charges; the energy is inversely proportional the the distance between the two charges (r_{ij}), ϵ_0 is the vacuum permittivity. The total potential energy derived from intra- and intermolecular electrostatic interactions is calculated as a sum of these pair-wise contributions [71].

The partial atomic charges are usually derived by a least square fit aiming to reproduce the precalculated molecular electrostatic potentials. The restrained electrostatic potential fit (RESP) to 6-31G*-derived electrostatic potentials for multiple conformations of various molecules was used in the Amber94 FF. The restrains are used to reduce the artificially high charges on buried non-polar atoms. Factor $^1/_{1.2}$ is was used to downscale the electrostatic interaction between 1-4 atoms [73]. It has been known that the fit to 6-31G* gas phase electrostatic potentials overestimates molecular polarities; this has been deemed desirable in explicit solvent simulations, as the commonly used TIP3P point charge water model has dipole moment about 20% larger than the corresponding gas phase water molecule. The charges derived from the 6-31G* gas phase fit are thus considered to implicitly contain some of the solvent polarisation effects [73, 74, 76].

For the Amber03 FF, partial atomic charges of amino acid atoms were derived by RESP fitting to the electrostatic potentials calculated at more precise B3LYP/cc-pVTZ/HF/6-31G** level in the presence of implicit solvent (ϵ = 4). The use of a non-unitary dielectric constant led to a different distribution of atomic charges compared to Amber94; this distribution is thought to be more similar to that occuring naturally in the condensed phase. Despite the partial charges being lower, slightly larger dipole moments were observed overall [76]. Unlike Amber99SB(-ILDN), Amber03 is often considered a distinct FF, due to the fundamentally different approach used for the derivation of partial charges [79].

The second term in the brackets in Equation 2.5 describes the dispersion and exchange-repulsion interactions. The latter, scaling here as r^{-12} is used to approximate the strong repulsion experienced by a pair of atoms as they are pulled close together. This repulsion stems from Pauli's exclusion principle, which prohibits for two identical fermions (in this case electrons) in a system to occupy the same region in space. This leads to a decrease of electron density in the interatomic region, which in turn results in the strong repulsion of the unshielded nuclei. The attractive part of the potential, scaling as r^{-6}, describes the London dispersion interactions, which result from the correlation of instantaneous electonic distributions between neighbouring atoms. Dispersion and exchange-repulsion interactions are part of the van der Waals forces, which also include forces between permanent electric multipoles (Keesom interactions, which are effectively included in the electrostatic term) and forces between permanent and induced multipoles (Debye

interactions) [71].

The particular functional form of Equation 2.5 is known as the Lennard-Jones 12-6 potential. It contains two parameters: σ_{ij}, which is the value of interatomic distance at which the potential passes through zero, and ϵ_{ij}, which is the value of the potential energy at the function minimum. The dependence of the attractive part on the interatomic distance is physically correct and can be derived from quantum Drude oscillators; its repulsive part is, however, too steep, as in reality the electron density decays exponentially. The popularity of the r^{-12}-scaling term is due to the ease with which it can be computed by squaring the dispersion component. Various other functional forms have been developed that model the repulsive part more realistically, for example the Buckingham potential:

$$v(r) = \epsilon \left[\frac{6}{\alpha - 6} e^{-\alpha\left(\frac{r}{r_m} - 1\right)} - \frac{\alpha}{\alpha - 6} \left(\frac{r_m}{r}\right)^6 \right] \tag{2.9}$$

This potential involves an extra parameter, α, setting which to a value around 15 causes the function to behave similarly to the Lennard-Jones potential near the minimum.

In Amber94, the Lennard-Jones parameters ϵ_{ij} and σ_{ij} were derived from Monte Carlo simulations of the condensed phase and empirically adjusted to reproduce the observed densities and enthalpies of vaporisation. Factor $^1/_2$ was used to scale down the 1-4 van der Waals interactions. This scaling is physically justified for two reasons. The first is that the r^{-12} repulsive term is too steep compared to the correct exponential potential and the associated error is the largest for the van der Waals interactions between pairs of atoms separated three covalent bonds. Second, the polarisation of the 1-4 atoms would lead to a decreased repulsion, but is not explicitly included [73].

In CHARMM22, the electrostatic and van der Waals terms had their parameters determined in a completely different way. The partial atomic charges were derived as to reproduce the 6-31G(d) interaction energies between the model compounds (for example, *N*-methylacetamide in the case of the parameters of protein backbone) and a TIP3P water molecule; initial charges were derived from the Mulliken population analysis [83] of the 6-31G(d) wave function and iteratively optimised. Dimer geometries utilising each polar site on the molecule were considered. The concept of groups of up to five atoms, in which the total charge is either 0 or ± 1 was utilised, allowing for a simplier application of the derived parameters to larger molecules. The van der Waals parameters were derived from explicit solvent simulations of the model compounds by fitting to experimental values of heats of vaporisation and molecular volumes. The partial charges were then readjusted to fit the interaction energies if necessary. In the CHARMM27 FF, the partial

atomic charges and Lennard-Jones parameters were rederived from fitting to HF/6-31G* *ab initio* interaction energy calculations between DNA nucleotide moieties and water and between Watson-Crick base pairs [78, 81]. The CHARMM22 and CHARMM27 FFs do not use any scaling of the 1-4 interactions [77, 78].

The properties of empirical methods that must be considered in their upcoming application are:

- A simple potential energy function is used, splitting the energy into well-interpretable contributions. This function does not account for many-body effects nor does it explicitly describe properties that depend on electron distribution (*i.e.*, polarisation effects).

- Terms for interactions between bonded atoms are in a form suitable for small deviations from equilibrium. The corresponding parameters (reference bond lengths or angles) were derived from crystal structures and *ab initio* calculations and are similar between different FFs.

- Parameters for non-bonded interactions were derived differently in each of the FFs. Effects of solvent were always considered in the parametrisation of van der Waals interactions and, in some cases, also in the derivation of atomic partial charges.

2.2.2 System partitioning and computational execution

The procedure atomising the interactions between proteins and DNA into the pairs of interacting residues described in Sections 2.1 and 2.1.1 resulted in the retrieval of amino acid – dNMP pairs. For multiple reasons, it was found desirable to get rid of the atoms constituing the protein backbone groups. First, the inclusion of C_α amide and carbonyl groups would introduce charged moieties into the molecule, greatly complicating the interpretation of the gas phase interaction energies (see Section 4.1). Second, each peptide bond group would have to be capped, creating intra- and intermolecular interactions that do not exist in nature. Finally, the properties of the atoms constituing the protein backbone are the same in each standard α-amino acid. Therefore, the binding motifs involving the peptide bond groups can hardly be viewed as being representative of some preferred interaction mode between a specific amino acid – DNA residue pair [53].

For these reasons, in each amino acid – dNMP dimer, the peptide bond carbonyl and amide groups of the amino acid were replaced with hydrogen atoms, in a process consistent with an earlier work on pairs of amino acids

by Berka *et al.* and other similar studies [44, 45, 84]. Each standard amino acid was therefore capped by a methyl group at the C_β atom; the result of this geometry culling is called the C_α representation of the amino acid. In the case of proline, only the carboxyl group was removed and a five-member heterocycle was retained [53]. In the rest of this study, the term amino acid in the context of interaction energy calculations refers to their respective C_α representations, unless stated otherwise.

To study the influence of the sugar-phosphate moiety on the interaction specificity, the set of dNMP – amino acid pairs (Table 2.4) was duplicated; for each dimer, one copy remained as it was, while from the other one the sugar-phosphate moiety was removed. Hydrogen atom added in place of the missing 2'-deoxyribose to the N9 atom of purine or N1 atom of pyrimidine bases [53].

Four sets of contacts resulted from the application of the above described culling stages to the set of amino acid – DNA nucleotide pairs:

- C_α representations of amino acids with DNA nucleotides
- C_α representations of amino acids with DNA bases
- C_α representations of amino acids with DNA nucleotides in which the amino acid side chain interacts directly with the DNA base moiety
- C_α representations of amino acids with DNA bases in which the amino acid side chain interacts directly with the DNA base moiety

As described in Section 2.1.2, a direct contact between the amino acid side chain and the DNA base was recognised when the distance between any two heavy atoms of the interacting partners was less than the sum of their van der Waals radii plus 1.0 Å. Furthermore, each set was studied at the four maximum sequence identity levels described in Section 2.1.1. The total number of structures in the first two sets after the redundancy criteria had been applied is summarised in Table 2.4; the populations of the last two are described in Table 2.5.

Due to the way nucleic acid residues are labelled in PDB structures, the extraction of the Nth DNA nucleotide resulted in the phosphate moieties lacking the O3' oxygen atom belonging to 2'-deoxyribose of the immediately preceeding (N-1)th residue. This atom was added to the structures as follows. A unit vector perpendicular to the plane containing atoms OP1, OP2 and O5' was defined by the normalised cross product of vectors corresponding to hypothetical bonds O5'-OP1 and O5'-OP2. This vector was then translated to the P atom and its length was scaled by factor 1.610; the O3' atom was

added at its end. This value was chosen because 1.610 Å is the reference length of the P-O3' bond in Amber94 FF [73].

As the dimers were extracted from X-ray structures only, no hydrogen atoms were originally present. This problem was remedied utilising a custom Chimera-1.8.1 script, which was used to add the hydrogen atoms to both the amino acid as well as DNA residues in all contacts. The program first examines the local environment around each heavy atom before any dissociable (acidic) hydrogens are added. Chimera-1.8.1 proved to be the only program capable of correctly assigning the hybridisation states, and therefore the correct number of hydrogens, to atoms in the C_α representations of amino acids [85].

The histidine side chains were set to have the dissociable hydrogen added to ϵ-N in all contacts, even if the local environment would suggest alternative protonation. Proline was modelled as a neutral tetrahydropyrrole and glycine as methane. Guanine and cytosine were represented by the dominant keto forms while adenine and thymine by the dominant amino forms. In purine bases, a hydrogen atom was added to the N9 nitrogen. Guanine was set to be protonated on N1 atom instead of N3 in all contacts, even of the local environment would lead to the hydrogen being added to N3. A single hydrogen atom was added to the phosphate group [53].

If the script failed to protonate any one of the interacting partners, the dimer was discarded. This happened in most cases due to hydrogen atoms not being added correctly to the 2'-deoxyribose moiety. The only exception when the dimer was repaired instead of being discarded after an incorrect protonation was in the abovementioned case of the alternative protonation of guanine, in which case the correction to N1 was possible due to the common frame of reference of the DNA bases. The number of discarded dimers was of the order of 10^1 and the number of (originally) misprotonated guanine bases was of the order of 10^2. The enumerations in Tables 2.5 and 2.5 provide the respective summaries after the incompletely protonated structured had already been discarded [53].

The parameters of the C_α representations of amino acids, of the isolated DNA bases, as well as of the dNMPs had to be added to the corresponding FF residue topology files. The atom types of the atoms not present in the original topologies were added based on chemical similarity as follows. Atom types of the added C_α hydrogen atoms were HC in Amber03 or AMBER99SB-ILDN and HA in CHARMM22 FFs. These atom types correspond to hydrogen atoms bonded to aliphatic carbon atoms without electron-withdrawing groups and general non-polar hydrogens, respectively. The atom type of the hydrogen atom added to proline nitrogen was H in Amber03 or AMBER99SB-ILDN, which is assigned to any hydrogen added

to a nitrogen atom. In CHARMM22, the atom type of the proline secondary amine hydrogen was also H; in this case, however, the parameters describe hydrogen atoms bonded to polar groups. The atom type of the hydrogen added to N9 atom in purine and N1 atom in pyrimidine bases in place of the C1' atom discarded with the sugar-phosphate moieties were (when applicable) H in Amber99 and HN2 in CHARMM27 FFs, respectively; the latter is the atom type of protons bound to DNA base ring nitrogen atoms. For the calculations involving DNA nucleotides, performed with the Amber99 FF, the atom type of the added phosphate group oxygen was OS, which the same as the one for the 2'-deoxyribose O3' atom; the atom type of the hydrogen atom attached to it was HO, which is used for any hydroxyl group hydrogen [53, 73–78].

Partial atomic charges are not fixed for individual atom types in the aforementioned FFs; atoms of the same atom type, even within a single molecule, can carry different partial charges depending on the parametrisation procedure (Section 2.2.1). The partial charges of the added atoms thus had to be manually entered after the topologies had been modified. For the added C_α hydrogens, the partial charges were symmetrically split between all added atoms so that the total charge of each amino acid residue was an integer: $+1.0$ for lysine and arginine, -1.0 for aspartate and glutamate and 0.0 for the remaining amino acids. Only the dominant form of each amino acid at $\mathrm{pH} = 7$ was thus considered; histidine was modelled as neutral, in agreement with the protonation described above. Four amino acids – glycine, alanine, valine and proline – obtained symmetry in the C_α representation that had not been previously present. No changes were made to the partial charges of the original atoms in alanine, valine and proline; however, in glycine, the partial charges were redistributed evenly between the four hydrogen atoms were to match the unchanged partial charge of carbon. This redistribution was performed because, otherwise, the properties of glycine would significantly differ depending on its orientation relative to its interacting partner; moreover, due to the desire to keep the molecule neutral, unphysically large charges would be attributed to two of the hydrogens of the original charges were to remain unchanged. All DNA bases were modelled as neutral, the partial charges of the added N9 (purine) or N1 (pyrimidine) hydrogen atoms were thus simply calculated to keep the overall charge null. In residues retaining the sugar-phosphate groups, the charge of the added O3' atom was the same as the charge of the O3' found in the 2'-deoxyribose moiety; the added phosphate hydroxyl group hydrogen had its partial charge assigned so that the charge of the whole dNMP residue was -1.0. In the Amber99 FF, four types of DNA residues are defined: those found at the 3' and 5' termini, those found within the DNA strands and free nucleosides. The parameters of

the residues found within DNA strands were used in this study, as the other choices lack some of the moieties needed (free nucleosides), or were found unbalanced (3' terminal residue) [53, 73, 74]. This choice of DNA residue type of not present in the CHARMM27 FF [78].

The interaction energy calculations were performed as follows. First, the geometry of the C_α representation of amino acid in dimer conformation with DNA base (or with dNMP) was taken and the positions of the hydrogen atoms were optimised in both partners while keeping the heavy atoms fixes. A single point energy was then calculated on this optimised complex ($E_{complex}$). The dimer in this minimised geometry was then split and a single point energy calculation was performed on each of the monomers (E_{1A}, E_{1B}; A and B are the constituing monomers). Afterwards, each of the monomers had the hydrogen atom positions optimised by itself. Heavy atoms were once again confined to their original positions. Single point energy calculation was the performed on each of the constituing monomers (E_{2A}, E_{2B}). The difference between the single point energy of the monomer after it had been isolated from the complex and after it was optimised by itself is the deformation energy of the monomer:

$$E_{defA} = E_{1A} - E_{2A} \tag{2.10}$$

$$E_{defB} = E_{1B} - E_{2B} \tag{2.11}$$

The deformation energy has a small positive value usually in the range of 0–2 kJ/mol (See Section 3.1, Figure 3.4). The difference between the single point energy of the optimised complex and the sum of the single point energies of the constituing monomers, after each had been optimised by itself, is the interaction energy:

$$E_{int,def} = E_{complex} - (E_{2A} + E_{2B}) \tag{2.12}$$

where the index $_{\mathrm{def}}$ indicates that the deformation energy is included in the result. This form of interaction energy is used in all further illustrations, except for the comparison with benchmark *ab initio* results. In this case, the deformation energy is ommited for consistency and the following formula is used:

$$E_{int} = E_{complex} - (E_{1A} + E_{1B}) \tag{2.13}$$

All MM interaction energy calculations were performed in the gas phase using program GROMACS-4.5.5 [86]. Double precision versions of the required tools had to be compiled and used together with conjugate gradient optimisation algorithm — otherwise, artifacts such as aliphatic chains remaining in the eclipsed confirmation resulted [53].

2.3 Electrostatic potentials

Electrostatic potential $\phi(\mathbf{r})$ at point $\mathbf{r}$ is the work done by the electric field in bringing a unitary positive electric charge from infinity to point $\mathbf{r}$; the potential energy of a point energy charge q at $\mathbf{r}$ is $\phi(\mathbf{r})q$. The electrostatic potential at point $\mathbf{r}$ is calculated as

$$\phi(\mathbf{r}) = \phi_{nucl}(\mathbf{r}) + \phi_{elec}(\mathbf{r}) \tag{2.14}$$

where

$$\phi_{nucl}(\mathbf{r}) = \sum_{A=1}^{M} \frac{Z_A}{|\mathbf{R}_A - \mathbf{r}|} \tag{2.15}$$

is the contribution from M nuclei located at $\mathbf{R}_A$ with charges Z_A, $A = \{1, 2, 3, \ldots, M\}$, and

$$\phi_{elec}(\mathbf{r}) = -\int \frac{\rho(\mathbf{r}')\mathrm{d}\mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|} \tag{2.16}$$

is the contribution from the electron density $\rho(\mathbf{r}')$ integrated over all points $\mathrm{d}\mathbf{r}'$ occupied by the electrons. Molecular electrostatic potential is the fuctional value of $\phi(\mathbf{r})$ calculated on a set of points defining the molecular surface [71, 87].

As long-range interactions between molecules are primarily of electric nature, molecular electrostatic potentials are frequently used to predict the binding properties of the investigated species. They are especially useful for the the assessment of interaction specificity in complexes of biomolecules, in which the recognition process is guided by non-covalent interactions. The surface electrostatic potential complementarity has been used to predict and design the interactions between two proteins [88, 89], as well as the formation of protein – nucleic acids complexes [90–92]. They have also been integral in the process of investigating the binding of small molecule ligands to their biomolecular targets [93, 94].

In order to understand the physical basis of the recognition of specific DNA sequence patterns by amino acids, the electrostatic potentials of each of the isolated DNA bases were first calculated and compared to each other. As noted in Section 2.2.1, the force field partial atomic charges were derived by fitting to the electrostatic potentials of molecular fragments [73–78]. Therefore, any results for complexes in which electrostatic interactions are dominant can be rationalised by and directly attributed to electrostatic potential observations. The effects of the addition of the sugar-phosphate moieties to the DNA bases were then studied, with focus on whether the electrostatic potential around the sugar-phosphate moiety is modified in a

sequence-dependent manner. Finally, perturbation of the electrostatic potentials of isolated DNA bases or dNMPs due to interaction with selected amino acids was investigated.

The geometries of the aforementioned species were extracted from contacts in the set of cluster representative. The only criterion was that the DNA base be in the *anti-* conformation. Electrostatic potentials were calculated in Gaussian 09 using cc-pVTZ basis sets and HF-level wave functions [95, 96]. Singlet states were considered for all molecules and molecular complexes. Subsequent visualisation was performed using Molden5.2.2 [97].

Chapter 3

Results

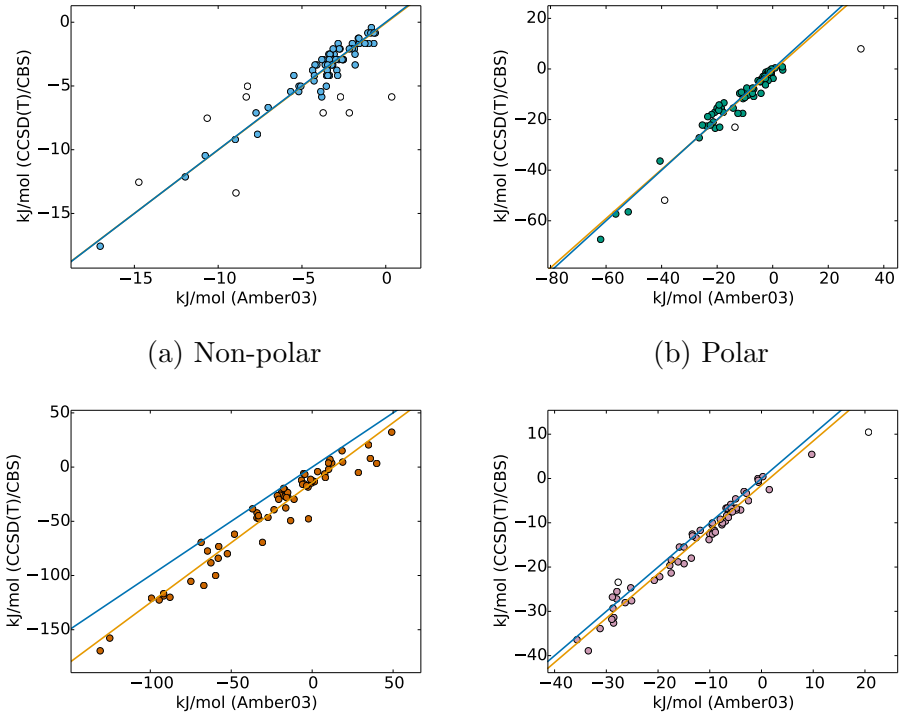
This chapter begins with the estimation of reliability of the introduced FFs for the calculation of interaction energies in amino acid – DNA residue pairs. This is done by a systematic comparison of the results provided by the respective empirical methods with a high-quality *ab inition* energies. In Section 3.2, a large scale study of binding preferences featuring tens of thousands of contacts from real protein – DNA complexes is performed. This section opens with the definition of how interaction specificity can be defined, following by application of the introduced criteria to progressively larger sets of structures. Finally, a qualitative analysis of electrostatic potentials around various residues and complexes is performed, in an attempt to provide sound physical basis for the observed preferences.

3.1 Evaluation of force field performance

As described in Section 2.2.1, the FFs used for the calculation of interaction energies had their partial charges and van der Waals parameters unanimously derived with solvent interactions in mind. Their application to the gas phase calculations is therefore questionable. The biological relevance of the gas phase approximation will be discussed in detail in Section 4.1. For now, the performance of the FFs will be investigated by comparison with the benchmark CCSD(T)/CBS energies.

The correlations of Amber03, Amber99SB-ILDN and CHARMM27 interaction energies with CCSD(T)/CBS results are shown in Figures 3.1–3.3. The benchmark interaction energies were calculated only for the contacts found in the set of cluster representatives. Furthermore, only the DNA base moiety in contact with the C_α representation of the amino acid were considered in these calculations. No deformation energy was included in either

the respective force field or benchmark results at this stage (*i.e.*, Equation 2.13 was used). This set counted a total of 272 DNA base – amino acid side chain pairs. These dimers were split into four groups based on the physico-chemical characteristics of each particular amino acid and analysed separately. Seventy-six pairs were found in the set of non-polar contacts featuring alanine, glycine, isoleucine, leucine, proline and valine; 69 comprised the polar set which features asparagine, cysteine, glutamine, methionine, serine and threonine; 64 involve charged amino acids arginine, aspartate, lysine and glutamate, and the remaining 63 were found in set involving aromatic amino acids histidine, phenylalanine, tryptophan and tyrosine. The out-



(a) Non-polar

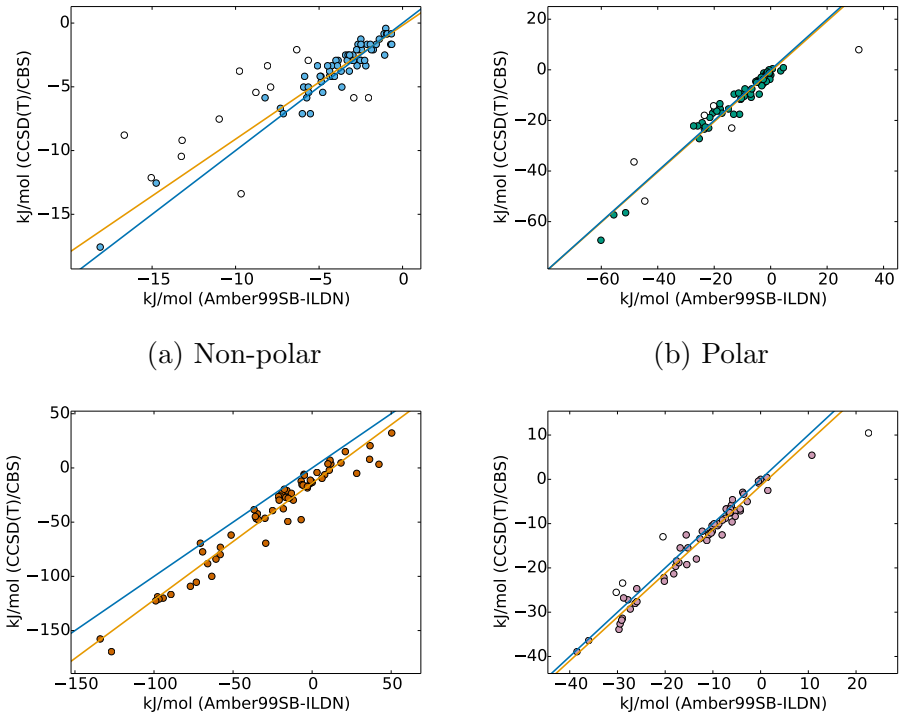
(b) Polar

(c) Charged

(d) Aromatic

Figure 3.1: Correlation of Amber03 and CCSD(T)/CBS interaction energies for contacts involving the respective physico-chemical types of amino acids. Deformation energy is not included. Outliers, marked as white, are defined in text and are not considered in the linear regression (orange line). Blue line corresponds to $y = x$.

liers marked by white circles in Figures 3.1–3.3 were defined as follows. The difference between the respective FF and CCSD(T)/CBS energy values was



(a) Non-polar

(b) Polar

(c) Charged

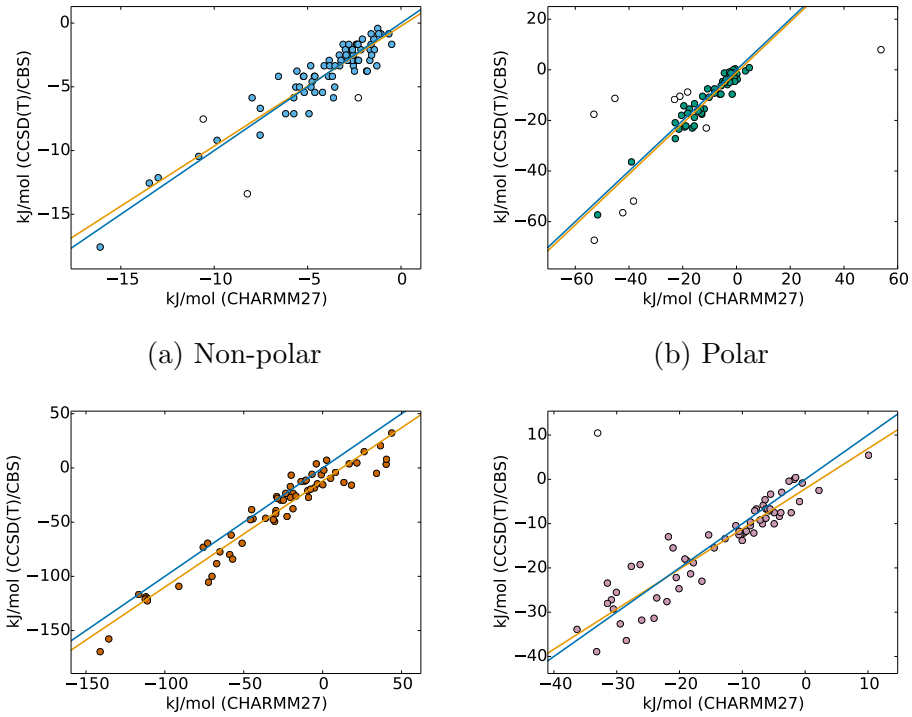
(d) Aromatic

Figure 3.2: Correlation of Amber99SB-ILDN and CCSD(T)/CBS interaction energies for contacts involving the respective physico-chemical types of amino acids. Details as in Figure 3.1.

calculated for each dimer ($\Delta x = x_{FF} - x_{CCSD(T)/CBS}$). The distribution of these differences was taken to be normal[1] and its third quartile ($Q3$), first quartile ($Q1$) and interquartile (IQR) values were calculated. When the particular interaction energy difference between the results of the two methods was greater than $Q3 + IQR$ or less than $Q1 - IQR$, the pair of energies x_{FF}, $x_{CCSD(T)/CBS}$ was marked as producing an outlier and was excluded from the linear regression analysis (represented by the orange line in Figures 3.1–3.3).

The mean signed errors (MSEs) associated with each of the FFs are summarised in Table 3.1. Positive values indicate that the respective FF interaction energies are underestimated compared to the benchmark. As already evident from Figures 3.1–3.3, dimers containing charged amino acids are sys-

[1]Shapiro-Wilk test confirms the assumption of normality to be valid at 95% confidence level only for the sets of contacts involving charged amino acids. In order to keep the analyses simple, the lack of normality in the other sets was overlooked and attributed to the limited sample sizes.



(a) Non-polar

(b) Polar

(c) Charged

(d) Aromatic

Figure 3.3: Correlation of CHARMM27 and CCSD(T)/CBS interaction energies for contacts involving the respective physico-chemical types of amino acids. Details as in Figure 3.1.

tematically destabilised in each of the used FFs. This can be attributed to the lack of explicit polarisation in the empirical potential energy function, which is more pronounced due to the calculations being performned in the gas phase. While the MSEs are positive for the other sets of amino acids as well, the standard deviations of the Δx (see previous paragraph) distributions (parentheses in Table 3.1) imply that the majority of data still lies near the blue lines in Figures 3.1–3.3, which have a unitary slope and indicate absolute correspondence of the two data vectors. This is especially true for the non-polar and polar sets, in which the standard deviations are an order of magnitude higher compared to the MSE. On the other hand, in the sets of contacts involving charged amino acids, the majority of data points lies in the vicinity of the regression line, which is shifted from the absolute correspondence by the respective MSE value close to the its particular standard deviation.

Coefficients of determination R^2 (Pearson's R squared) describe how

	Non-polar (kJ/mol)	Polar (kJ/mol)	Charged (kJ/mol)	Aromatic (kJ/mol)
Amber03	0.1 (0.7)	0.5 (2.7)	17.0 (11.6)	1.5 (1.7)
Amber99SB-ILDN	-0.2 (0.8)	0.6 (2.6)	16.0 (10.9)	1.3 (1.7)
CHARMM27	0.0 (1.0)	1.1 (2.8)	11.3 (11.7)	0.9 (3.8)

Table 3.1: MSEs of the tested force fields for contacts involving the respective physico-chemical types of amino acids. CCSD(T)/CBS energies were taken as reference "true" results. Standard deviations of the differences between the two results are indicated in parentheses. Positive MSE values indicate destabilisation by the particular FF. No deformation energies were included. Outliers were not considered in the calculations.

many percent of variance in one data vector (FF energies) is explained by the variance in another data set (benchmark results). They are summarised for the respective FFs and contacts featuring the particular physico-chemical sets of amino acids in Tables 3.2 and 3.3; the latter includes outliers in the calculations.

	Non-polar	Polar	Charged	Aromatic
Amber03	0.94	0.96	0.94	0.97
Amber99SB-ILDN	0.92	0.96	0.95	0.97
CHARMM27	0.89	0.93	0.93	0.87

Table 3.2: Coefficients of determination R^2 between the tested force field and CCSD(T)/CBS energies for contacts involving the respective physico-chemical types of amino acids. No deformation energies were included. Outliers were not considered in the calculations.

The agreement between the respective FF and benchmark energies is very good if outliers are excluded from the comparison (Table 3.2). The R^2 values are sometimes as high as 97%. Little difference is seen between the particular FFs, except for the decreased average performance of CHARMM27 when dealing with complexes involving aromatic amino acids.

Unfortunately, no benchmark energies are available in any of the subsequent analyses and hence the definition of outliers is not possible. Moreover, the large-scale data are strongly non-normal (see below), further complicating the statistical treatment. The correlation coefficients from Table 3.3, in which the outliers were included, are, therefore, more appropriate indicators of the computational performance[2] of the respective FFs for the applications

[2]Which does not, as stated, imply anything about the biological relevance of the con-

	Non-polar	Polar	Charged	Aromatic
Amber03	0.79	0.81	0.94	0.96
Amber99SB-ILDN	0.77	0.81	0.95	0.95
CHARMM27	0.84	0.68	0.93	0.65

Table 3.3: Coefficients of determination R^2 between the tested force field and CCSD(T)/CBS energies for contacts involving the respective physico-chemical types of amino acids. No deformation energies were considered. Outliers were included in the calculations.

to follow. The inclusion of the outliers makes the difference between the Amber class and CHARMM27 FFs much more pronounced. The average performance of CHARMM27 deteriorates mostly when dealing with outliers in the set of contacts involving aromatic amino acids, while both Amber FFs are affected the most when pairs featuring non-polar amino acids are treated. The reliability of all force fields gets worse when dealing with contacts featuring polar amino acids, for the the charged complexes are left unchanged as, of course, no outliers were identified in this group.

To illustrate the treatment of the real sets of contacts further, Table 3.4 summarises the MSEs found for the respective FFs and amino acid types when the outliers were included in the calculations. No changes in the trends

	Non-polar (kJ/mol)	Polar (kJ/mol)	Charged (kJ/mol)	Aromatic (kJ/mol)
Amber03	0.2 (1.5)	2.3 (10.6)	17.0 (11.6)	1.6 (2.1)
Amber99SB-ILDN	-0.7 (1.9)	1.9 (10.7)	16.0 (10.9)	1.1 (2.7)
CHARMM27	0.1 (1.3)	2.0 (12.7)	11.3 (11.7)	0.2 (6.7)

Table 3.4: MSEs of the tested FFs for contacts involving the respective physico-chemical types of amino acids. Outliers were included in the calculations. Other details as in Table 3.1.

of over- or understabilising the interaction energies were observed. In comparison with Table 3.1, in which the outliers were excluded, the standard deviations of the Δx distributions rose in all sets in which outliers were found. This effect is most dramatic when investigating the treatnent of complexes with polar amino acids, in which the standard deviation is almost four times higher.

To investigate where do these discrepancies originate, three interaction energy outliers were chosen from the comparison of Amber99SB-ILDN and

sidered gas phase and pair-wise approximations (see Sections 4.1 and 4.2).

CCSD(T)/CBS results (Figure 3.2) and examined in detail. These examples represent the most extreme case of discrepancy in complexes involving the respective types of amino acids. The interaction and deformation energies of these pairs were compared with benchmark values and summarised in Table 3.5. The reason the CCSD(T)/CBS deformation energies are not

	IE (kJ/mol)		E_{def} (kJ/mol)		IE + E_{def} (kJ/mol)	
	FF	benchmark	FF	benchmark	FF	benchmark
Non-polar	-16.7	-8.8	2.7	1.6	-14.0	-7.2
Polar	31.2	8.0	1.9	3.7	33.1	11.7
Aromatic	22.6	10.5	4.4	6.3	27.0	16.8

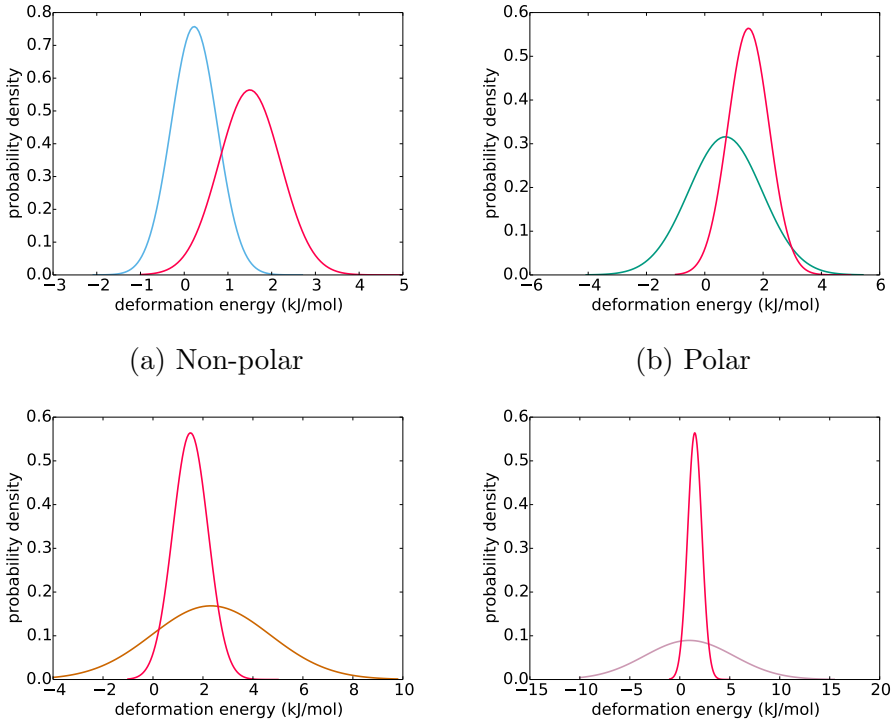
Table 3.5: Interaction and deformation energies of the complexes causing the largest discrepancies between the Amber99SB-ILDN and CCSD(T)/CBS results. Benchmark deformation energies were obtained from DFT-D/B3LYP-D3/def2-TZVPP (see text).

routinely included in the previous analyses is that the *ab initio* geometry optimisation was performed at the DFT-D/B3LYP-D3/def2-TZVPP level and they are therefore not available (Jiří Hostaš, personal communication). Hence, the benchmark deformation energies mentioned in this section originate from this method.

The distributions of deformation energies for the complexes involving the respective physico-chemical types of amino acid are very strongly normal (as proven by Shapiro-Wilk test, results not shown) and are illustrated in Figure 3.4 (negative values are, of course, extrapolated). Comparing this figure with the discrepancy-causing results shown in Table 3.5, it becomes visible that the complexes is which the FFs fail to describe the correct interaction energy are those in which either the FF or benchmark deformation energies are close to the high end of the deformation energy spectrum. These large deformation energies indicate that the respective pairs are highly strained in the dimer configuration. The treatment of these non-equilibrium geometries, therefore, seems to be flawed at the FF level. While the rationale for the disagreememt with benchmark results remains the same for Amber03 and CHARMM27 FFs, the parameter nuances of these FFs lead to the different observed average performance (Table 3.3 and 3.4).

One can summarise the results of this section in the following points:

- The average performance of Amber03, Amber99SB-ILDN and CHARMM27 FFs is in exceptionally good agreement with the benchmark method as long as the amino acid – DNA base pairs are reasonably close



(a) Non-polar (b) Polar

(c) Charged (d) Aromatic

Figure 3.4: Amber99SB-ILDN deformation energy distributions for complexes involving the respective physico-chemical types of amino acids (blue, green, vermillion, pink). DFT-D/B3LYP-D3/def2-TZVPP deformation energies are shown in crimson. Negative values are extrapolated.

to the equilibrium geometries (*i.e.*, the interaction energy difference outliers are excluded).

- The reliability of the respective FFs differs when dealing with dimers in which the approximations involved in the empirical treatment or the particular potential energy functional form are not applicable.

- When it is not possible to identify the contacts providing outlier interaction energy differences, the average performance of the aforementioned FFs deteriorates quickly. The decrease in agreement with the benchmark results depends on each particular FF.

3.2 Large-scale binding preferences

Having explored the reliability of the empirical methods on the limited set of cluster representatives, their application to calculating the interaction energies of the tens of thousands complexes (Table 2.4) forming the amino acid side chain distributions became desirable. As described in Section 2.1, the distribution of a certain amino acid around a particular DNA base is the set of all respective DNA base – amino acid pairs extracted from protein – DNA complexes available at the four sequence identity criteria stated in Section 2.1.1. Mirror sets, containing dNMPs instead of the isolated DNA bases, as well as subsets of both of these containing only direct contacts with the DNA base (Table 2.5), were introduced. Quantitative analysis of these sets will be the subject of this section.

To investigate the relative energetics of various interaction modes in as concise a manner as possible, an interaction energy profile was created for each of the $4 \times 20 \times 4 \times 4 = 1,280$ (see previous paragraph) distributions using Amber99SB-ILDN parameter set results [74, 80]. The particular FF was chosen for three reasons. First, it provided balanced results in comparison with the benchmark interaction energies when the outliers were included (Section 3.1) [76]. Second, it derives the partial charges for both protein and DNA residues using the same approach, unlike Amber03 (Section 2.2.1). Finally, it serves as a basis for the parmbsc0 FF, which is very commonly used for the simulations of protein – DNA complexes [98]. The interaction energy profiles were constructed as follows. First, a histogram of the interaction energies of all amino acid – DNA base (dNMP) complexes in each respective distribution was created. The number of bins was calculated by applying the Freedman-Diaconis formula [99]. The interaction energy profile of each cluster in that distribution was then created by making a histogram of the cluster members' interaction energies, respecting the bin boundaries calculated for the distribution. The respective histograms were replaced by an interpolated function and overlayed on top of each other. This procedure enabled relatively quick evaluation of the energetics of the binding modes the clusters represent and their comparison with the those found in the bulk of the distributions (see below).

The interaction energy profiles of a fraction of these distributions, involving only the pairs in which the amino acids interacts directly with the DNA bases, extracted from complexes originating from applying 100% maximum sequence identity criterium, have already been published [53] and are available online at `http://bioinfo.uochb.cas.cz/projects/pdna-iea/`. All profiles are provided in the digital supplementary materials.

The amino acid clusters in each distribution represent geometrically con-

served interaction modes. This conservation does not, however, automatically imply a role in the direct sequence readout mechanism. For example, contacts featuring single hydrogen bond donor or acceptor groups are naturally sterically constrained because of the directional requirements of hydrogen bonds and are therefore prone to being found in clusters. Yet, single hydrogen bonds are not sufficient to distinguish between individual DNA bases. Based solely on the dimer geometry, the possibilities of specific base recognition by a single amino acid by the means of a unique hydrogen bond donor/acceptor group pattern are therefore limited to the few pairs featuring bidentate hydrogen bonds [16].

It is desirable to augment this definition of specificity by explicitly considering the interaction energies of the respective dimer conformations. The following points summarise my line of reasoning for what is to be considered an amino acid – DNA base (dNMP[3]) pair geometry significant for the sequence recognition process:

- The orientation of the amino acid relative to the DNA base (dNMP) must be found within one of the geometrical clusters. This condition implies that the respective interaction mode is utilised by many protein – DNA complexes. Therefore, it is not bound to be functional only under some unique local environment of a single protein family.

- The cluster to which the pair belongs must correspond to the most favourable arrangement of the two partners. In other words, it must have the (signed) lowest interaction energy found for that particular amino acid – DNA base pair.

- The peak corresponding to the low-lying cluster must be separated from the bulk of the distribution. This condition ensures that once the mutual orientation of the partners providing specific sequence recognition is achieved, the energetic gain is such that a reversal to an unspecific geometry is improbable.

- No other contacts other than those belonging to the distinct low-lying cluster are to be present within its interaction energy range. This criterion has two consequences. First, it enables the identification of specificity-determining pair geometries based on the respective interaction energies. Second, it implies that all pairs within that particular

[3]Although the sugar-phosphate moiety is chemically the same in all dNMPs, its local environment can be modified a depending on the base and therefore provide different response to the approaching interacting partner. If a cluster is localised around the sugar-phosphate group and is conserved at various maximum sequence identity levels in a base-dependent manner, it can be viewed as containing important sequence-specific contacts.

interaction energy range are highly sterically specific, as they could have been identified as forming a cluster.

- The previous criteria specify energetically distinct geometries within distributions. For an amino acid A to uniquely distinguish between individual DNA bases, the interaction energies found in pairs from the identified distinct low-lying cluster must also be lower (signed) than those provided by any contacts of that particular amino acid with any other base type. In other words, the stabilisation of the complex A – B, where B is the recognised DNA base, adopting a conformation falling to the distinct cluster, must be greater than the interaction energy found for any pair of A with any other base type. This distinction is to be made for each of the nucleotide edges — major groove, minor groove and backbone — separately, as it may be possible for an amino acid to uniquely distinguish between different bases in each these regions.

Only when meeting all these criteria can the coupling between energetic and geometrical aspects of specificity be achieved. A clear drawback of this interaction energy profile-based specificity definition is that only sufficiently represented motifs will be detected. If this analysis is to be considered complete, it must be assumed that all amino acid preferences towards DNA bases can already be detected in the binding modes realised in the currently available structures of protein – DNA complexes. The deficiencies regarding the inadequate treatment of interactions with larger DNA residue blocks as well as other many-body effects are discussed in Section 4.2. The following paragraphs will describe the application of the described rules to the distributions and the localisation of the distinct low-lying clusters.

3.2.1 Interactions directed at the DNA bases

To begin, only the complexes in which the amino acid interacts at least partially with the DNA base were considered in the distributions. As described in Section 2.1.2, this is a subset of amino acid – dNMP or amino acid – DNA base contacts in which the distance between between any pair of heavy atoms belonging to the amino acid and the base moiety is less than the sum of the atoms' van der Waals' radii plus 1.0 Å. The exclusion of the contacts directed solely at the 2'-deoxyribose or phosphate groups simplifies the initial analyses of interaction specificity.

Table 3.6 summarises for which amino acid – DNA base type pairs were the distinct low-lying clusters found in the respective interaction energy profiles. Table 3.7 presents the dNMP – amino acid pairs the interaction energy profiles of which contain distinct clusters. The used set of structures was

created by augmenting the previous one with the sugar-phosphate moiety at each DNA base. The influence of the charged sugar-phosphate group on the base-directed specific interactions can be examined by comparing these two tables.

Identity	30%	90%	95%	100%
Adenine	N,Q	N,Q	N,Q	N,Q
Cytosine				
Guanine	R	R	R	R
Thymine				

Table 3.6: DNA base – amino acid pairs whose interaction energy profiles contain distinct low-lying clusters at the respective maximum sequence identity levels. Only the complexes in which the amino acid is in direct contact with the DNA base were used for construction of the interaction energy profiles.

Identity	30%	90%	95%	100%
dAMP	Q	N,Q	N,Q	N
dCMP				
dGMP				
TMP		Y	Y	Y

Table 3.7: dNMP – amino acid pairs whose interaction energy profiles contain distinct low-lying clusters at the respective maximum sequence identity levels. Only the complexes in which the amino acid is in direct contact with the DNA base were used for construction of the interaction energy profiles.

It can be seen that contacts of adenine with asparagine and glutamine are of particular importance at all sequence identity levels. The criteria defining specificity-determining cluster geometries are not broken in the presence of the sugar-phosphate moiety in these pairs. The interaction energy profiles for the adenine (dAMP) – asparagine and adenine (dAMP) – glutamine pairs are shown in Figures A.3 and A.4, respectively.

Figure A.5 compares the interaction energy profile of the dAMP – asparagine pair with complexes involving other DNA nucleotides. Note the higher total number of dAMP – asparagine pairs observed. The lack of clusters or any significant populations of contacts within the interaction energy range of the distinct low-lying cluster is evident in contacts featuring any dNMP other than dAMP (Figures A.5b – A.5d). Therefore, it can be concluded that, for the adenine – asparagine pair, there exists a conformation

of the partners that, when assumed, is capable of uniquely distinguishing between individual bases based on interaction energy criteria. The same behaviour is observed for the adenine (dAMP) – glutamine pairs.

Atomic-level examination of the distinct low-lying cluster reveals that the respective contacts correspond to a bidentate hydrogen bond geometry (Figure A.6). The asparagine (glutamine) side chain amide group serves as a hydrogen bond donor and acceptor for the adenine N6 amino group donor and N7 acceptor atoms. This interaction mode is realised in the major groove of the DNA double helix, where the bidentate hydrogen bond donor and acceptor pattern is capable of uniquely distinguishing between adenine and other bases.

Energetically distinct clusters were also found in the interaction energy profiles of the guanine – arginine pairs (Table 3.6). Unlike the contacts with adenine, the addition of the sugar-phosphate moiety to the guanine base shears the distinction between the peak corresponding to the distinct cluster conformations and the bulk of the distributions (Figures A.7c and A.7d). This is related to the gas phase approximation, which results in a strong long-range attraction between the guanidino and phosphate groups. As the electrostatic term becomes dominant, the interaction energy for a particular pair becomes increasingly proportional to the distance between the two charged groups, with other features of the complexes playing secondary roles. The gas phase approximation will be discussed in detail in Section 4.1.

Figure A.8 shows the interaction energy profiles of the complexes of arginine with all DNA bases. It can be seen that not only are the distinct low-lying clusters not found in any distribution other than that of the guanine – arginine pair, but also all the other profiles terminate before the range of interaction energies of the specificity-determing cluster is achieved. In other words, the geometries of the guanine – arginine pairs contributing to the low-lying cluster are more stabilising than any other geometry of arginine with any other base. Note the higher total population of this pair compared to contacts with the other bases.

The respective interaction energy profiles deserve a special commentary in this case. For the adenine – asparagine (glutamine) pairs, all requirements on the possibly specificity-determining clusters were met. However, when the guanine – arginine pairs are considered, a considerable envelope of non-cluster contacts covers the lowest lying cluster (Figures A.7a and A.7b). The answer to why the actual cluster contacts were still regarded as specific despite failing to meet the unique interaction energy criterium lies in the way the clusters were constructed. The identification of clusters based on the RMSD of atom positions is inappropriate when two or more conformations of an amino acid provide the same pattern of hydrogen bond donors and

acceptors. In that case, the different conformations will not fall into the same (or any) cluster, because the RMSD between the positions of identical atoms is too large. If the two symmetry-related conformations provide the same interaction energies, a fraction of the contacts within that particular interaction energy range will not be identified as members of a cluster, despite interacting with the same functional group.

The guanine – arginine pairs fits precisely into the above described category. The guanidino group provides two hydrogen bond donors which can be manifested by either the two terminal or a terminal and ϵ-N atoms. Moreover, a mirror image of each binding mode can be generated, which provides the same pattern of hydrogen bond donors. The geometries of the pairs found in the low-lying cluster feature two hydrogen bonds only between the terminal nitrogen atoms of the arginine guanidino group and the O6 and N7 acceptors of guanine (Figure A.9a). One of the alternative isoenergetic geometries is shown in Figure A.9b. Contacts like these do not contribute to the cluster population, but are still regarded as specific at the one amino acid – one DNA base correspondence level [16, 30, 53]. These interactions take place in the major groove, where guanine is the only base providing two hydrogen bond acceptor atoms, making geometrical distinction from other bases possible by means of two amino acid hydrogen bond donor groups.

While the sugar-phosphate moiety diminishes the specificity of the arginine – guanine pair, an opposite effect is observed in the interaction energy profiles of the thymine – tyrosine distributions (Figure A.10). Although no distinct clusters were identified when only the bases were considered, the addition of the charged group shifted the interaction energies of members of a single cluster into the most favourable region of the interaction energy spectrum. The phosphate group does, therefore, contribute to the clustering of tyrosine relative to thymine.

The interaction energy profiles of tyrosine pairs with all four dNMPs are compared in Figure A.11. It can be seen that although only in the thymine – tyrosine distribution does the low-lying cluster appear, geometries providing the same stabilisation energies are also available for complexes with the other bases. Moreover, the populations of the distribution with each of the DNA bases are very similar. This is in contrast to the distributions of adenine – asparagine (glutamine) and guanine – arginine, in which the specific pairs had populations significantly higher than the other dimers (see Figures A.5 and A.8).

A tyrosine – thymidine 5'-monophosphate (TMP) pair chosen from the distinct cluster is illustrated in Figure A.12. It features a single hydrogen bond between the tyrosine hydroxyl group and the phosphate moiety. No specific interactions with the base are observed. A visual examination of

complexes of tyrosine with the other dNMPs reveals that each distribution contains pairs providing the same interaction energies adapting an orientation of the partners similar to Figure A.12. However, the complexes with the other DNA residues also feature geometries not present in contacts with TMP, as illustrated in Figure A.13 on an example dAMP – tyrosine pair. These orientations provide the same interaction energies as the complex with TMP, although the contacts do not form clusters in the respective distributions.

Combining the results presented in the last two paragraphs, it is clear that although the geometry shown in Figure A.12 is the most favourable one for the TMP – tyrosine pair, it does not distinguish between individual DNA base types based on the interaction energies alone. The reason why the cluster is observed only for this pair may be its greater chance appearance in the relatively few protein – DNA complexes available. The populations of dimers with other bases featuring this conformation may simply not yet be sufficient to form clusters.

Unfortunately, it is clear that the definition of and search for the distinct clusters is somewhat subjective. In order to not miss any important contacts, Table 3.8 summarises the amino acid – base pair distributions in which some low-lying were found, although they did not fully meet some of the required criteria. Table 3.9 contains the same summary but considers pairs of amino acids with dNMPs instead. In both of these analyses, only the dimers in which the amino acid interacts directly with the DNA base moiety are included. The mentioned diminishing of energetic specificity of the low-lying

Identity	30%	90%	95%	100%
Adenine				R,K,T
Cytosine	D	D	D	N,W
Guanine	D	D	D	D
Thymine	R	R	R	R,N,S,T

Table 3.8: DNA base – amino acid pairs whose interaction energy profiles also contain low-lying clusters at the respective sequence identity levels. These clusters are less distinct than those in Figure 3.6 (see text). Only the complexes in which the amino acid is in direct contact with the DNA base were used for construction of the interaction energy profiles.

cluster in the guanine – arginine distribution after the inclusion of the sugar-phosphate moiety is evident from Table 3.9, as the contacts excluded when strict requirements on the distinct clusters were applied (Table 3.7) appear here. Few new clusters have been identified, featuring non-polar (isoleucine, tryptophan), polar (serine, threonine) and even charged (lysine, aspartate)

Identity	30%	90%	95%	100%
dAMP	N			Q
dCMP				N,D,I,K
dGMP	R	R	R	R,W
TMP	S	S	S	N,S,T

Table 3.9: The dNMP – amino acid pairs whose interaction energy profiles also contain low-lying clusters at the respective sequence identity levels. These clusters are less distinct than those in Table 3.7 (see text). Only the complexes in which the amino acid is in direct contact with the DNA base were used for construction of the interaction energy profiles.

species. Some of these will be briefly investigated in the following paragraphs, while others will be investigated in detail later in this section.

Figure A.14 shows the interaction energy profile and a representative of the low-lying cluster found in the distribution of adenine – threonine pairs. This complex features a single hydrogen bond between the threonine hydroxyl group donor and adenine N3 acceptor atoms. The interaction takes place in the minor groove. Comparing the respective interaction energy profile to, for example, that of the adenine – asparagine pairs (Figure A.3b), the less distinctive characteristics of the adenine – threonine cluster become apparent.

Another cluster pronounced only when the isolated DNA bases are considered features thymine – arginine pairs. As shown in Figure A.15, this contact is realised in the minor groove and involves two hydrogen bonds between the donor guanidino group and acceptor O2 atom of the base. The corresponding interaction energy profile reveals that stabilisation energies provided by the members of this cluster can partially be provided by some non-cluster geometries.

Deoxycytidine 5'-monophosphate (dCMP) – isoleucine pairs form another distribution containing a low-lying cluster (Figure A.16). This was surprising, as non-polar amino acids were not expected to provide highly stabilising directional interactions. Adding to the mystery, the cluster appears detectably pronounced only when the sugar-phosphate moiety is included. The pair geometry shows the amino acid approaching the nucleotide from the minor groove and forming van der Waals contacts with the 2'-deoxyribose and base moieties. It is unlikely that this binding motif would not include any of the neighbouring base steps.

The last two clusters reviewed involve contacts of TMP with serine or threonine. Both of these pairs feature similar interaction energies, which can, however, also be partially provided by other, non-cluster members of the

respective distributions (Figures A.17a and A.17b). Very similar geometries are assumed by both pairs, featuring a single hydrogen bond between the amino acid hydroxyl group and one of the phosphate oxygen atoms. van der Waals contacts with the major groove-exposed atoms are also present. The distinctive energetical characteristics of both of these clusters are clearly dependent on the presence of the phosphate group.

Large stabilisation energies for the respective distribution had already been already observed in threonine – adenine pairs forming a low-lying cluster (Figure A.14). These contacts involved a hydrogen bond interaction with the DNA base moiety in the minor groove. The arrangement was energetically less favourable compared to the threonine – TMP complexes. The stabilisation energy of the latter can be, however, expected to be exaggerated in the gas phase. Both motifs were therefore considered as potentially significant. As each of the motifs is realised in a different groove of the DNA double helix, it is unlikely that both are utilised at the same time by a single DNA-binding protein.

Having investigated the extraordinary cases, it is now desirable to statistically analyse the interaction energies at large. Table 3.10 presents the median values of the interaction energies for complexes with DNA bases involving the respective physico-chemical types of amino acids. This separation was done due to the significantly different interaction energy ranges found in the individual groups. Median values were chosen because of the presence of outliers, which are hard to treat properly, as the distributions strongly deviate from normality. Median absolute deviations[4] (MADs) were used to characterise the spread of the data. It can be seen from the distribution median and MAD

	IE, 30% (kJ/mol)	IE, 90% (kJ/mol)	IE, 95% (kJ/mol)	IE, 100% (kJ/mol)
Non-polar	−4.2 (1.8)	−4.2 (1.7)	−4.2 (1.7)	−4.2 (1.6)
Polar	−8.3 (6.9)	−8.4 (7.0)	−8.4 (6.9)	−8.4 (6.9)
Charged (+)	−24.9 (29.7)	−24.0 (29.3)	−23.8 (29.4)	−24.0 (29.3)
Charged (−)	−25.9 (32.3)	−28.8 (31.6)	−29.1 (31.3)	−31.6 (30.5)
Aromatic	−8.9 (5.7)	−8.7 (5.4)	−8.8 (5.5)	−9.2 (5.6)

Table 3.10: Median interaction energies found in amino acid – DNA base complexes involving the respective physico-chemical groups of amino acids. MADs in parentheses. Only the pairs involving direct contacts with the DNA base were considered. Various maximum sequence identity levels are shown.

[4]These are calculated by finding the median of the absolute diffences between each data point and the median of the data set.

values that the vast majority of interactions are favourable when only the DNA bases are considered. This is especially true for the contacts involving non-polar amino acids, in which next to no postive (repulsive) interaction energies were found. Moreover, it can be seen that the interaction energy ranges of the distinct low-lying clusters identified at this level — pairs of adenine – asparagine (~ -40 kJ/mol), adenine – glutamine (~ -50 kJ/mol) and guanine – arginine (~ -120 kJ/mol) — are all significantly more stabilising than an average interaction involving the respective physico-chemical group of amino acids. The same remains true for all the pairs which had the low-lying cluster identified with some reservations (Table 3.8).

Table 3.11 contains the large-scale interaction energy energy characteristics of amino acid – dNMP dimers in which the amino acid interactions directly with the base moiety. As expected, the interactions involving charged

	IE, 30% (kJ/mol)	IE, 90% (kJ/mol)	IE, 95% (kJ/mol)	IE, 100% (kJ/mol)
Non-polar	−7.1 (3.7)	−7.2 (3.7)	−7.3 (3.7)	−7.4 (3.6)
Polar	−16.4 (12.6)	−16.7 (12.8)	−16.7 (12.8)	−16.3 (12.3)
Charged (+)	−198.1 (56.2)	−198.4 (56.2)	−198.1 (56.5)	−197.4 (55.3)
Charged (−)	127.4 (41.0)	122.7 (40.8)	122.7 (41.2)	121.8 (40.9)
Aromatic	−15.7 (10.5)	−15.3 (9.9)	−15.5 (10.0)	−15.1 (9.4)

Table 3.11: Median interaction energies found in amino acid – dNMP complexes involving the respective physico-chemical groups of amino acids. MADs in parentheses. Only the pairs involving direct contacts with the DNA base were considered. Various maximum sequence identity levels are shown.

amino acids became significantly more stabilising or repulsive depending on the charge the amino acid carries. The contacts featuring non-polar, polar or aromatic amino acids have all shown increased stabilisation, although the corresponding MADs grew as well. The discovered cluster in the TMP – tyrosine distribution (Figure A.10c), as well as the less clearly defined distinct clusters (Table 3.9), once again provide interaction energies significantly more stabilising than an average contact featuring an aromatic amino acid. It must be added, however, that the changes observed after the addition of the charged sugar-phosphate moiety have likely been greatly exaggerated by the calculations being performed in the gas phase (Section 4.1).

3.2.2 Interactions with the DNA backbone

So far, only the contacts in which the amino acid interacts at least partially with the DNA base were considered for analyses. This allowed me to eval-

uate the relative interaction energies of various interaction modes featuring the sequence-determining moieties. Although the influence of the phosphate group was studied in these complexes, no comparison of these base-directed contacts with the energetics of pairs involving interactions with the sugar-phosphate backbone was made. Moreover, all sequence-dependent differences in amino acid affinities for the sugar-phosphate moieties were naturally ignored. The study of these two previously omitted features will be the subject of the following paragraphs.

Table 3.12 summarises the amino acid – DNA base pair distributions in which distinct low-lying clusters were found when all contacts were considered in the construction of the interaction energy profiles. Any preferences

Identity	30%	90%	95%	100%
Adenine	N,Q	N,Q	N,Q	N,Q,K,M
Cytosine				N,K
Guanine	R,D	R,D	R,D	R,D
Thymine	R	R	R	

Table 3.12: DNA base – amino acid pairs whose interaction energy profiles contain distinct low-lying clusters at the respective maximum sequence identity levels. All complexes in the distributions were used for construction of the interaction energy profiles.

that appear here and do not appear in Tables 3.6 and 3.8 show pairs in which the interaction is geometrically and energetically distinct, but is not primarily directed at the DNA base. Notable additions include the pairs adenine – lysine, cytosine – asparagine, cytosine – lysine and guanine – aspartate (Figure A.18). It must be added, however, that most of these distinct clusters are observed only when the relatively benign redundancy criterion of removing only identical protein – DNA complexes is applied. This suggests that these motifs are likely not universally shared, as the proteins from which they were extracted contain homologous domains.

Interestingly, the interaction of aspartate with guanine is conserved at all redundancy levels. Figure A.19 shows the comparison of the interaction energy profiles of aspartate pairs with all DNA bases. One can see that no complex with any other base[5] provides interaction energies between -100 kJ/mol and -150 kJ/mol like the contacts in the distinct cluster found in

[5]Although the blue cluster in the cytosine – aspartate distribution (Figure A.19b) seems promising, it loses about a half of its population once any stricter redundancy reduction criterion is applied. The envelope of isoenergetic contacts covering it then becomes much more pronounced.

the guanine – aspartate distribution.

Atomic-level detail of a member from this guanine – aspartate cluster brings surprising results. The amino acid interacts with guanine *via* a bidentate hydrogen bond between the acceptor aspartate terminal carboxyl group and donor N1 and N2 amino group atoms of the base (Figure A.20a). These atoms are, however, usually involved in the Watson-Crick pairing between guanine and cytosine, with which the interaction with aspartate interferes. On the other hand, the slightly less stabilising cytosine – aspartate contacts found in the blue cluster in Figure A.19b feature a single hydrogen bond between the aspartate carboxyl group and the C4 amino group of the base (Figure A.20b. This interaction occurs in the naturally accessible major groove.

It is interesting that six guanine – aspartate dimers featuring the geometry shown in Figure A.20a were found even after the most strict redundancy reduction criterion had been applied, that is, the sequence identity of any two proteins the set from which the dimers were obtained was less than 30%. The PDB IDs of the protein – DNA complexes from which the guanine – asparate pairs were extracted are 1jb7 (twice), 1omh, 1po6, 1xjv and 3zh2. While the proteins involved differ significantly with respect to their structure, they unanimously interact with either telomeric, aptameric or otherwise deformed DNA molecules. While probably not involved in routine sequence recognition, the highly stabilising interaction shown in Figure A.20a does contribute and can even be crucial for the recognition of few non-canonical forms of DNA.

The inclusion of all contacts in the distributions in the construction of the interaction energy profiles had little impact on the specific clusters of adenine – asparagine (glutamine) and guanine – arginine pairs. In general, only an increase in the frequency of the interaction energies provided by the bulk of the distributions was observed (Figure A.21). This implies that the gemetry featuring the bidentate hydrogen bond binding motif is more stable than any possible non-specific arrangement of the interacting partners.

In accord with the previous treatment of contacts directed at the DNA bases, Table 3.13 presents amino acid – DNA base pair distributions in the interaction energy profiles of which were identified low-lying clusters which fail to meet some of the distinctive criteria. The only notable additions that do not appear in Tables 3.6, 3.8 and 3.12 are the clusters in guanine – histidine and thymine – histidine distributions.

Let's now briefly discuss the slightly distinct cytosine – tryptophan and guanine – tryptophan pairs. Interestingly, the corresponding interaction energy profiles were both classified as containing bluntly defined low-lying clusters in Tables 3.8 and 3.9, respectively. The geometries of these pairs and

Identity	30%	90%	95%	100%
Adenine				D
Cytosine				D,I,W
Guanine	L	H,L	H,L	L,M,W
Thymine		K	K	R,N,H,K,P,S,T

Table 3.13: DNA base – amino acid pairs whose interaction energy profiles also contain low-lying clusters at the respective sequence identity levels. These clusters are less distinct than those in Table 3.12 (see text). All complexes were used for construction of the interaction energy profiles.

their interaction energy distributions are shown in Figures A.22a and A.22b. While the stabilisation energies provided by both pairs are similar, the respective interaction modes differ significantly. The contact with cytosine features a stacked conformation in which the plane of the tryptophan indole ring is parallel to the plane of the base atoms. On the other hand, the complex with guanine involves a single hydrogen bond between the indole group nitrogen donor and the base N3 acceptor atoms. This interaction is realised in the minor groove and remains significant even after the sugar-phosphate moieties are included (Table 3.9).

One might wonder what is the relevance of calculating the interaction energies of amino acid – base dimers in which the amino acid does not naturally interact directly with the base. These systems are, of course, artificial, as in the protein – DNA complexes from which the dimers were extracted the interaction is realised through the sugar-phosphate. In MM methods, the non-bonded potential energy terms are calculated by summing between all pairs of atoms in the complex (Section 2.2.1). The contribution of the non-covalent interactions between the base moiety and the amino acid, regardless of the pair geometry, to the potential energy of the complex is the same as with the sugar-phosphate attached due to this pair-wise nature of the empirical methods. It was shown in the case of the deoxyguanosine 5'-monophosphate (dGMP) – arginine pairs that the phosphate group blunts the distinctive characteristics of the low-lying cluster observed when only the base is considered (Figure A.7). The pair containing a charged amino acid, this blurring of the interaction preferences was related to the gas-phase approximation. Therefore, I rationalise that the separation of the base moiety from the complex, albeit artificial, is suited for the estimation of binding preferences of individual amino acids to the respective DNA residues, as it rids one of the artifacts associated with the calculations being performed *in vacuo*. This separation does, however, eliminate any contacts in which the

sugar-phosphate moiety contributes to or even carries the specificity markers guiding the preferential amino acid binding. The explicit inclusion of these contacts will be the finale of the interaction energy profile studies.

The dNMP – amino acid pair distributions in which distinct low-lying clusters were found are summarised in Table 3.14. Comparing these results

Identity	30%	90%	95%	100%
dAMP	Q	N,Q	N,Q	N,Q
dCMP	Q	Q	Q	N,Q,H,I
dGMP				
TMP	Q	Q,Y	Q,Y	Q,Y

Table 3.14: The dNMP – amino acid pairs whose interaction energy profiles contain distinct low-lying clusters at the respective maximum sequence identity levels. All complexes in the distributions were used for construction of the interaction energy profiles.

with Table 3.12, in which in only the DNA bases were considered, one can see similar differences to those observed by comparing the results of the calculations performed only on base-directed contacts (Tables 3.6 and 3.7). Namely, the identified clusters of charged amino acids manifested by the adenine – lysine, cytosine – lysine, guanine – arginine, guanine – aspartate and thymine – arginine pairs lose their distinctive characteristics as the interaction energies for the respective contacts within these clusters blend in with the bulk of the distributions.

One of the clusters that appears after the inclusion of the sugar-phosphate is for the already described TMP – tyrosine pair (Figure A.12). As shown in Figure A.11, the amino acid in this geometry does not interact with any base-dependent groups, and so this conformation might correspond simply to the most favourable arrangement of tyrosine in complex with any base (dNMP).

A completely new distinct clusters appears in the distributions of glutamine – TMP or dCMP pairs. Figure A.23 shows how do the respective interaction energy profiles of the thymine – glutamine pairs change after the sugar-phosphate moieties are included. The distributions involving cytosine – glutamine pairs are very similar, both as to the interaction energy ranges and profile characteristics. It can be seen that from Figure A.23 that the interaction energy of the members of this cluster changes by about 100 kJ/mol once the sugar-phosphate moiety is included.

Figure A.24 compares the interaction energy profiles of glutamine – dAMP, dCMP, dGMP and TMP pairs. There is an energetic distinction between

the low-lying cluster found in the dAMP – glutamine and TMP (dCMP) – glutamine distributions. While the interaction energy range of the distinct cluster is between −60 and −80 kJ/mol in the dAMP – glutamine distribution, both bounds are shifted by about 20 kJ/mol towards more negative values in the TMP (dCMP) – glutamine pairs. This larger stabilisation is attributed to the calculations being performed in the gas phase. I do in no way challenge the specificity of the adenine – glutamine interaction, noting that the respective binding motifs are realised in different regiones around the DNA double helix. The adenine – glutamine hydrogen bonding takes place in the major groove, while the interactions with purine base nucleotides are directed at the DNA backbone. Therefore, the significance of both motifs for the direct sequence readout mechanism is possibly depending on the orientation of the interacting partners: the purine bases can be recognised from the backbone edge of the double helix, while adenine is distinguished in the major groove.

Atomic-level investigation of the representatives of these distinct clusters reveals that the only interaction involved is a single hydrogen bond between the terminal donor amide group of the amino acid and a phosphate group oxygen acceptor (Figure A.25). These contacts do not involve any contribution from the base moiety, unlike the abovementioned complexes of TMP – tyrosine complexes (Figure A.12). In fact, the exclusion of the sugar-phosphate moieties makes these pairs slightly repulsive (Figure A.23). It is interesting that these geometries are conserved even after the most strict redundancy reduction criterion was applied. Moreover, the described interaction mode only involves contacts featuring pyrimidine bases. An attempt to explain these nuances by the means of electrostatic potentials will be made in Section 3.3.

For completeness, Table 3.15 summarises the dNMP – amino acid pairs in the distributions of which were identified low-lying clusters which did not fully meet the required distinctive criteria. As in the previous cases, most

Identity	30%	90%	95%	100%
dAMP	N			H,K,T
dCMP				
dGMP	R,D,L	R,D,L	D	N,D,S,T
TMP	S	S	S	S,T

Table 3.15: The dNMP – amino acid pairs whose interaction energy profiles also contain low-lying clusters at the respective sequence identity levels. These clusters are less distinct than those in Table 3.14 (see text). All complexes were used for construction of the interaction energy profiles.

contacts forming bluntly-defined low-lying clusters are found in complexes involving either charged or polar amino acids. In the former, the interaction energy-based distinction is blurred by the dominant electrostatic term which does not depend on the geometry of the pair. On the other hand, many complexes which feature a single hydrogen bond involving polar amino acids are not directional enough to form highly populated clusters. Therefore, envelopes of other isoenergetic contacts exist in the respective interaction energy profiles.

In accord with my previous statistical treatment of the interaction energy distributions at large, Tables 3.16 and 3.17 contain the summaries of interaction energy characteristics of all amino acid – DNA base and amino acid – dNMP pairs, respectively. Let's first compare the former to Table 3.10, in

	IE, 30% (kJ/mol)	IE, 90% (kJ/mol)	IE, 95% (kJ/mol)	IE, 100% (kJ/mol)
Non-polar	−1.7 (1.5)	−1.8 (1.6)	−1.7 (1.6)	−1.3 (1.3)
Polar	−2.3 (3.4)	−2.1 (3.2)	−2.1 (3.2)	−1.6 (2.7)
Charged (+)	0.0 (14.9)	0.4 (14.4)	0.4 (14.4)	0.6 (14.0)
Charged (−)	−13.7 (13.6)	−13.0 (13.3)	−12.8 (13.1)	−12.3 (12.4)
Aromatic	−3.8 (3.9)	−3.5 (3.6)	−3.3 (3.6)	−2.9 (3.3)

Table 3.16: Median interaction energies found in amino acid – DNA base pairs involving the respective physico-chemical groups of amino acids. MADs in parentheses. Various maximum sequence identity levels are shown.

which the results for only a subset of pairs involving direct amino acid – DNA base are shown. Although most interactions remain attractive, an unanimous shift of the interaction energies towards lesser stability is observed. This suggests that the added contacts, despite experiencing attraction from the base, require the sugar-phosphate moiety to achieve stabilisation energies similar to those provided by the base-directed interactions. This is especially true for the contacts involving positively charged amino acids, which provide mixed attractive/repulsive interaction energies in the absence of the negatively charged sugar-phosphate. Once again, all contacts found in distinct low-lying clusters, as well as the investigated less pronounced cytosine and guanine – tryptophan pairs, provide interaction energies significantly more stabilising than the observed average values (Figure 3.12).

After the sugar-phosphate moieties were added to the DNA bases (Table 3.17), the average interaction energies of contacts involving non-polar, polar or aromatic amino acids returned aproximately to the values observed in Table 3.10. This confirms my hypothesis that, for some of these pairs, the

influence of the charged phosphate group is necessary to provide interaction energies comparable to those observed for base-directed contacts. The non-polar or aromatic amino acids rarely display specificity towards any single DNA base, unless they feature a hydrogen bond. However, the contacts featuring these amino acids are also almost universally favourable. Therefore, it seems rational to conclude that these pairs are can be viewed as serving as a stabilising glue which can be used without strict directional requirements. The complexes involving charged amino acids do, of course, become

	IE, 30% (kJ/mol)	IE, 90% (kJ/mol)	IE, 95% (kJ/mol)	IE, 100% (kJ/mol)
Non-polar	−7.1 (3.3)	−7.1 (3.2)	−7.2 (3.2)	−7.2 (3.1)
Polar	−16.7 (11.7)	−16.9 (11.9)	−17.0 (11.8)	−17.2 (11.6)
Charged (+)	−210.0 (51.9)	−211.4 (52.2)	−210.9 (52.1)	−212.3 (51.0)
Charged (−)	154.5 (38.2)	154.8 (40.4)	155.5 (40.8)	157.4 (40.6)
Aromatic	−14.3 (9.2)	−13.8 (8.7)	−13.8 (8.7)	−13.9 (8.5)

Table 3.17: Median interaction energies found in amino acid – dNMP complexes involving the respective physico-chemical groups of amino acids. MADs in parentheses. Various maximum sequence identity levels are shown.

significantly more stabilising/repulsive after the sugar-phosphate moieties are included. This change is slightly larger than when only base-directed contacts are considered (Table 3.11). Once again, the distinct low-lying clusters identified in the TMP – glutamine and dCMP – glutamine pairs' interaction energy profiles (Figures A.24b and A.24d) provide interaction energies significantly more stabilising than is typical for complexes involving polar amino acids.

The results related to this section can be summarised as follows:

- The members of some clusters provide stabilisation energies significantly larger than any other geometry available for that particular amino acid – DNA base type combination. In other words, the most favourable arrangement of the interacting partners is sometimes realised only within a very narrow window of mutual orientations.

- The complexes forming these energetically distinct clusters may contain amino acids of any physico-chemical type, although polar residues capable of forming hydrogen bonds are preferred. This is related to the more directional nature of the hydrogen bond compared to other non-covalent interaction motifs.

- When a unique one-to-one geometrical correspondence based on hydrogen bond donor/acceptor groups complementarity is possible between the interacting partners, the respective amino acid residues form distinct clusters. The interaction energies provided by such arrangements help distinguish between individual bases, favorising the geometry enabling the recognition. This applies to adenine – asparagine, adenine – glutamine and guanine – arginine pairs.

- More importantly, coupling of geometrical preference and energetical favourability was observed even for pairs where no unambiguous recognition by the hydrogen bond donor/acceptor pattern is possible. The vocabulary of contacts contributing to the direct readout mechanism of sequence recognition was thus expanded by using interaction energy-derived specificity criteria. This preference towards certain base types was found notably in cytosine glutamine, thymine – glutamine and thymine – tyrosine pairs, and to a lesser extent also in adenine – threonine, cytosine – aspartate, cytosine – isoleucine, thymine – arginine, thymine – serine and thymine – threonine contacts.

- The energetically distinct orientations may feature interactions with any combination of the base, 2'-deoxyribose and sugar-phosphate moieties. The respective binding can take place in both minor and major grooves, feature a stacking interaction, or even involve solely the sugar-phosphate backbone.

- The phosphate group contributes to the stability of some energetically distinct interactions directed at the base moiety. On the other hand, interactions involving only contacts with the sugar-phosphate displaying preference towards certain nucleotides were recognised.

- Non-specific contacts provide similar interaction energies whether they involve interactions with the sugar-phosphate group or the base. This is especially true for the pairs involving non-polar amino acids, which are almost universally stabilising, but do not show preference for any nucleotide.

- A specific interaction utilised in the recognition of highly deformed DNA regions was observed. Therefore, the interaction energy-based specificity criteria are robust enough to recognise both generic as well as niche binding motifs.

- It must be added, however, that most clusters do not appear to have any role in sequence recognition. Moreover, some of the distinct clus-

ters appear only when relatively benign sequence identity criteria are applied. The distinction between functionally significant contacts and redundant entries is difficult and sometimes subjective. It is unlikely that a simple algorithm would be able to distinguish between geometries significant for sequence recognition and non-specific contacts.

3.3 Electrostatic potentials

So far, the attraction between the amino acids and DNA residues forming specific contacts was largely described by a limited array of binding motifs: hydrogen bonds, van der Waals contacts and stacking interactions. The presence of the functional groups involved in these motifs is, however, a binary attribute, and as such it can hardly justify the observed preferences, which are likely guided by more subtle differences between individual residues. In particular, complementarity of hydrogen bond donor and acceptor groups can not account for the preference of some amino acids towards certain DNA bases acquired upon the addition of the sugar-phosphate moiety. Likewise, the many contacts featuring only a single hydrogen bond or an interaction with the sugar-phosphate can hardly be completely rationalised by this crude scheme. For these reasons, qualitative examination of the electrostatic potentials was used to obtain information about the more subtle differences between individual residues.

Initially, the electrostatic potential maps of isolated DNA bases were studied. The sugar-phosphate groups were then added in order to answer two questions introduced in the previous paragraphs, *i.e.*, how do the properties of the bases change in the presence of the charged group and whether some base-specific aspects are propagated onto the DNA backbone moieties. The electrostatic potentials of the DNA residues are shown in Figures A.26–A.33. Contour value of 0.01 and the same color coding are used in all illustrations, allowing direct qualitative comparison. Different views are provided, focusing on either the top of the bases or on their Watson-Crick edges, enabling the examination of different binding modes.

In isolated adenine, positive potential values are found on multiple atoms of the base: C2, C8, N9, and on the hydrogens of the C6 amino group (Figures A.26 and A.27). Negative values are found on the N1, N3 and N7 atoms. In the presence of the sugar-phosphate group, however, the negative charge is distributed thorough the DNA base, significantly lowering the potential values at the N1, N3, C5, N7, N9 and C6 amino group nitrogen atoms, with the only positive values remaining on the hydrogens of the amino group.

In the case of isolated cytosine, extensive area of positive electrostatic

potential is found surrounding the N1, C5, C6 and C4 amino group atoms (Figures A.28 and A.29). Negative values are found on the O2 keto group and N3 atoms. These become much more pronounced in the presence of the phosphate group. Transfer of the electron density onto the DNA base is apparent, with the potential dropping to negative values on the N1, C5 and C4 amino group nitrogen atoms. Positive electrostatic potential remains only on the hydrogen atoms of the amino group.

Isolated guanine base displays positive electrostatic potential values on the N1, C8, N9 and C2 amino group atoms, while negative values are found on the N3, O6 keto group and N7 atoms (Figures A.30 and A.31). The latter are the primary acceptors of the excess charge distributed thorough the base in the presence of the phosphate group. In addition, the potential turns negative on the C5 and N9 atoms, with positive values remaining only on the N1 and C2 amino group atoms. The decrease of potential on the amino group nitrogen is less pronounced than in the adenine and cytosine nucleotides, likely because of the presence of other atoms with high electron affinities.

Finally, isolated thymine base displays positive electrostatic potential surrounding all atoms except for the O2 and O4 keto groups (Figures A.32 and A.33). When the excess charge of the phosphate group is distributed thorough the DNA base, these oxygen atoms are its primary acceptors. However, decrease of potential in the presence of the sugar-phosphate moiety is observed on each atom, with positive value remaining only on the N3 hydrogen.

These results show that the electrostatic potentials of the DNA bases are strongly modulated by the presence of the phosphate group in their respective nucleotide forms. The empirical methods do not, of course, treat effects such as electron density transfer, which were responsible for the change of electric properties observed in these calculations. Instead, the phosphate group affects the interaction energies only through the distance-dependent non-covalent term (Section 2.2.1). These quantum effects are, however, implicitly responsible for the observed clustering of contacts, as they contribute to the preferences of amino acids towards DNA residues in natural conditions. The force field parameters used for the atoms of isolated DNA bases were the same as those for the bases moieties of nucleotides. It is, however, difficult to estimate where do they stand between the two extremes shown by the electrostatic potential maps.

Some of the observed phenomena can be rationalised by the results of these calculations. For example, a large decrease of electrostatic potential at the Hoogsteen edge of guanine in the presence of the sugar-phosphate (Figures A.30 and A.31) can be related to the delocalisation of the excess

electrons, which in turn enables multiple interactions outside the sterical constrains of the cluster to provide similar stabilisation energies. This is especially apparent when the electrostatic term is dominant, for example, in guanine – arginine pairs (see Figure A.7). The negative electrostatic potential of some otherwise neutral atoms of the base in the presence of the charged phosphate group can also contribute to the stability of pairs which display sequence-dependent preferences only in the presence of the phosphate. However, I also found that the base moiety does not affect the electrostatic potential around the sugar-phosphate in any significant way. Therefore, the origins of the apparent specificity found in some pairs interacting solely with the DNA backbone (Figure A.25) remain unexplained. The displayed preferences can, of course, be an artifact of the gas phase approximation and the limited size of the data set (see Sections 3.2 and 4.1 for related discussion).

Having observed the electrostatic potential maps of isolated DNA bases and nucleotides, it is possible to examine how do they predispose individual DNA residues towards interactions with specific amino acids. Moreover, one can look at how do the electric properties of the DNA bases change upon the interaction. These questions will be answered by a qualitative examination of the electrostatic potentials of amino acid – DNA residue complexes representing the contacts identified in Section 3.2 as possibly significant for direct DNA sequence recognition. These contacts are not presented in the order of assumed importance; some representatives are even chosed from clusters with less distinct characteristics.

Figures A.34–A.42 feature pairs which involve hydrogen bonds. Contacts of adenine with asparagine and threonine are shown in Figures A.34 and A.35, respectively. These interactions have subtle effects on the electronic distribution of the bases, with slight changes apparent only in the region between the atoms forming the respective hydrogen bonds.

Figure A.36 shows an interaction between aspartate and cytosine. Comparing the electrostatic potential around the base to that shown Figures A.28 and A.29, one immediately notices a significant transfer of negative charge onto the atoms of the DNA base. This effect is even more pronounced than that introduced by the presence of the phosphate group.

Figure A.37 illustrates a bidentate hydrogen bond between dGMP and arginine. Significant changes are observed in the electric properties of both partners upon interaction. The electrostatic potential surrounding free arginine is positive everywhere at the chosen contour value (not shown). Transfer of charge onto the amino acid is thus observed. Comparing the potential of the bound DNA nucleotide to that of its free form (Figures A.30 and A.31), it can be seen that this transfer happens mostly from the base moiety, which is significantly more positive when the complex is formed.

Bidentate hydrogen bond between guanine and aspartate is shown in Figure A.38. This motif was recognised in Section 3.2.2 as being involved in binding highly deformed DNA conformations. Similarities with the cytosine – aspartate complex (Figure A.36) are seen, such as the significant negative charge transfer onto the atoms of the DNA base.

Figure A.39 shows an interaction between guanine and tryptophan. Very little change in electric properties in either of the partners is observed outside the region of the hydrogen bond.

Contacts of TMP with serine and threonine are shown in Figures A.40 and A.41, respectively. Both feature a hydrogen bond between the hydroxyl group of the amino acid and one of the phosphate oxygen atoms. Little other information is provided by the electrostatic potential maps. In the complex featuring threonine, additional stabilisation is possible due to van der Waals contacts between the terminal methyl group of the amino acid and the C5 methyl group of the DNA base.

Figure A.42 shows a contact of TMP with tyrosine. It was said in Section 3.2.1 that similar geometries are also adapted by pairs of tyrosine with the other DNA residues, although no clusters were found in the respective distributions. A possible explanation for the observed preferences of the amino acid towards thymine is revealed by the electrostatic potential maps. The C5 methyl group of the base, in addition to sterically keeping the amino acid in place, allows additional dispersion interaction with the hydroxyl group due to its neutral charge and favourable orientation. This stabilisation is not present in contacts with the other bases, in which negatively charged (adenine, guanine) or no (cytosine) atoms are available for interaction.

Finally, in Figures A.43 and A.44 are shown contacts of dCMP with isoleucine and cytosine base with tryptophan. These complexes do not feature any hydrogen bonds. The latter pair adapts stacking conformation and does no display any significant electrostatic potential changes of either partner upon interaction. On the other hand, the contact with isoleucing reveals polarisation of the amino acid due to the presence of the charged phosphate group. Areas displaying enhanced potential potential are then in contact with negatively charged atoms of the base, suggesting an electrostatic interaction mechanism.

The results of this section can be summarised as follows:

- Electric properties of DNA bases are heavily modified in the presence of the phosphate group, with the negative charge distributed thorough the atoms of the base.

- In pairs of neutral amino acids with DNA bases, little to no transfer of charge transfer takes place between the molecules. This is valid

for contacts featuring hydrogen bonds, as well as for those bound by dispersion interactions.

- When hydrogen bonding is involved in the complex, and at least one of the interacting molecules is charged, large transfer of electron density occurs. This charge transfer can, in turn, change the electric properties of either molecule.

- Promimity of the charged phosphate group can affect the electronic distribution of a non-polar molecule, leading to induced electric moment which can aid stabilise the complex.

- As the attraction of specific amino acid – DNA residue pairs contacts is in most cases caused by the electrostatic complementarity of the molecules, electrostatic potentials provide a robust tool to explore the physical basis and effects of the binding. In some cases, they can also be used to rationalise the stabilisation of complexes with non-polar residues.

Chapter 4

Discussion

This chapter begins with the description of the approximations that were made over the course of interaction energy calculations. The intrinsic deficiencies of the empirical methods are reviewed in general and illustrated on the already heavily discussed amino acid – DNA residue pairs introduced in Section 3.2. These deficiencies are considered from a biological point of view; comparison of the computational performance of these methods with high-quality *ab initio* calculations was already made in Section 3.1 and will not be repeated. Various limitations of the pair-wise approximation are then discussed and few notes on true many-body effects are provided. A comparison with the few studies performed on similar systems is made, highlighting the conclusions of each.

4.1 Gas phase approximation

The most severe approximation used was, without doubt, the use of the gas phase. The lack of solvent molecules affects not only the calculation of interaction energies, but also strips some amino acid – DNA residue pairs utilising bridging water molecules in their respective interactions motifs of their biological relevance. Luscombe *et al.* found in a study performed on 129 non-redundant structures of protein – DNA complexes that as much as one fifth of all contacts is realised through a water molecule. Seventy percent of these water molecules were involved in interaction with the DNA backbone [30]. The use of water molecules in sequence recognition process has been documented, for example, in the case of the *trp* repressor/operator complex. In this structure, three well-ordered water molecules are crucial for the specific recognition of the DNA sequence. Each of these molecules is involved in three or four hydrogen bonds, enabling contacts between several

amino acids and DNA base steps [19]. However, Luscombe *et al.* found that most water molecules at the interface were found to be involved in only one or two hydrogen bonds, suggesting their role as non-specific modulators of the stability of the complex [30].

Regardless of whether they enable specific contacts or serve as generic fillers, all water molecules were removed from the protein – DNA complexes before any dimers were extracted. Althought each amino acid – DNA residue pair can be traced back to the structure from which it was extracted to investigate the presence of water molecules, the role of the solvent in the recognition process can not be retroactively predicted from the interaction energy profiles alone. Another limitation stems from the fact that the number of water molecules visible in a crystal structure is highly dependent on the resolution to which the structure was solved. For example, when Davey *et al.* solved the crystal structure of a nucleosome particle in complex with DNA to a resolution 1.9 Å, over 2,500 more water molecules were found compared to the same structure solved at 2.6 Å resolution. In addition to bridging more distant elements, these water molecules were found important in minor groove interactions, where they were involved in the binding of arginine side chains. The solvent was also found to reduce the sequence-dependency of nucleosome positioning [4]. This example illustrates that, in the range of resolutions considered in this study, substantial differences in the number of solvent-mediated interactions that can be observed exist. Therefore, even if one would consider all water molecules in the considered structures in the calculation of interaction energies, the full biological picture would not be reproduced due to the insufficient refinement of some complexes.

While the solvent-mediated interactions are featured in only a fraction of the contacts, the use of the gas phase affects all interaction energy calculations, as it leads to more or less severe overestimation of electrostatic interactions. This is most pronounced in the complexes of charged amino acids with dNMPs, which provide extremely stabilising or destabilising interaction energies depending on the charge of the amino acid involved. The fact that all amino acid – DNA residue dimers were extracted from real protein – DNA complexes puts the relevance of the large repulsive energies into question, as one does not expect any of the few contacts that are available for sequence recognition to significantly destabilise the complex. One of the specific pairs detected featured an interaction between guanine and a negatively charged aspartate (Figure A.20a) that was shown to be important for sequence recognition of some highly strained regions of DNA (Section 3.2.2). This interaction becomes repulsive when the sugar-phosphate moiety is added to the base, even though the binding motif involves hydrogen bonds with atoms at the Watson-Crick edge of the base and the distance between the amino acid

and the DNA backbone is therefore the greatest possible. On the other hand, the presence of the phosphate group was found to limit the preference of the guanine – arginine towards the conformation featuring a bidentate hydrogen bond with atoms in the major groove (Figures A.7c and A.7d), which is a canonical binding motif [16]. As described in Section 3.2.1, when the interaction energy of a particular pair is dominated by the electrostatic term, the stability of various binding motifs becomes less dependent on the geometry of the pairs. In complexes which do not contain charged amino acids, the effects of the calculations being performed *in vacuo* are still visible, although they are less predictable and more dependent on the relative orientation of the interacting partners.

The treatment of solvent effects effectively by the means of implicit solvation is one subject that I would like to research further. While the water-mediated sequence recognition would still be overlooked, one could get a much more relevant picture of interactions between charged residues. The application of implicit solvation models to nucleic acids is more difficult compared to proteins due to the complicated electrostatic properties of the polyelectrolyte molecule. Nevertheless, implicit solvent molecular dynamics simulations of free DNA [100,101] as well as protein – DNA complexes [69,102,103] have been showing promising results. Although improving in their ability to reproduce explicit solvent and experimental results, the optimal choice of atomic parameters[1] and solvation model is still not clear [103].

4.2 Pair-wise approximation and many-body effects

As this study was focused on finding binding preferences at the one amino acid – one DNA base correspondence level, contacts spanning multiple base steps or featuring interactions with both DNA residues in a base pair were not explicitly treated. More specifically, when an amino acid interacts with both Nth and $(N+1)$th bases in the DNA strand, two dimers were included in the data set, featuring the same amino acid in complex with the respective bases. If one were to look for specificity towards these larger blocks in as exhaustive a manner as it was done for individual DNA bases, that is to investigate the interaction energy profiles of all combinations of DNA base types in complex with all amino acids, it would very quickly become apparent that, given the amount of possible combinations, the contacts provided by

[1]The generalised Born parameters for some DNA residue atoms were not present at all in default GROMACS-4.5.5 [86] ports of the force fields used in this study.

currently available protein – DNA structures would not suffice. This may be concluded from the fact that there were some distributions containing a very limited amount of amino acid – DNA residue pairs already in this one-to-one correspondence study. One could use MD simulations to study the preferences towards these larger blocks, although intricate constrains limiting the conformational space to sample only the subset of orientations accessible when the block is part of a DNA double helix would have to be used. This restriction is not needed in this study, as all pairs were already extracted from real structures. The MD simulation could, however, be used to investigate the relative interaction energies even for binding motifs sparsely represented in the available structures. The systematic evaluation of contacts spanning multiple base steps was not performed by Luscombe *et al.* due to the limited size of the data set at the time [30].

In my opinion, the omission of the explicit treatment of contacts with multiple DNA bases is less severe compared to the lack of the water molecules in solvent-mediated interactions. Although each limitation can be viewed as a removal of one partner from a three body system, the bridging solvent molecule can be used either to extend a hydrogen bond over several ångströms, or even to bind two hydrogen bond acceptor groups, potentially changing repulsive interaction into a favourable one. On the other hand, as I do not expect repulsive interactions to be found in the limited space of the protein – DNA interface, I find it unlikely that an interaction with multiple bases would be necessary to stabilise an otherwise unfavourable binding motif. Moreover, the interactions spanning multiple base steps are partially included in this study, although they are atomised into multiple amino acid – DNA residue pairs. The preferences of amino acids towards larger DNA blocks could, in principle, be reconstructed from the available interaction energies, although the data set size would be, as already mentioned, far from that necessary for a sound statistical analysis involving all combinations of base pairs.

Benos *et al.* have shown that the assumption of additivity of individual amino acid – mononucleotide interactions is a reasonable approximation in the search of DNA binding sites [104]. Although their analysis was based on statistical knowledge-based potentials, it can be easily extended to preferences detected by interaction energy calculations, the results of which were not available at the time of their study. The main argument against the use of non-additive models in the prediction of protein binding sites was, however, the insufficient size of the data set at that time, which would not allow the investigation of certain binding motifs at all [104].

If one were to investigate the non-additivity of individual amino acid – DNA residue interactions by also varying the protein side, the number of

possible combinations that would have to be considered would quickly go beyond the limit of what can be observed in the currently available structures of protein – DNA complexes. While there are a total of $4 \times 4 = 16$ possible dinucleotides, the number dipeptides is $20 \times 20 = 400$, resulting in $16 \times 400 = 6{,}400$ dinucleotide – dipeptide pairs that would have to be sufficiently represented to allow any reasonable comparison. As there are currently over 20,000 amino acid – DNA residue pairs available in complexes of non-identical proteins (Table 2.4), about three representatives of each dinucleotide – dipeptide dimer would be found if each amino acid and nucleotide had the same probability of being found at the interface. In reality, the majority of combinations would not be found at all, given the already sparse population of some distributions in this study. The use of molecular dynamics simulations in the theoretical study of these larger blocks does, therefore, seems to be necessary.

The above described modulations of interaction specificity by a third molecule (water in solvent-mediated contacts and a second nucleotide in binding motifs spanning multiple base steps) must, of course, be distinguished from the true many-body effects introduced by the inclusion of additional species in the supermolecular system. The perturbing molecule affects the electronic distribution of the interacting molecules, leading to dispersion and induction effects. These changes of molecular properties are not explicitly treated in most empirical methods [82], including any of the tested force fields [73–78].

Moving to larger protein and DNA blocks interacting in water environment, the role of entropy, which was completely ignored in this study, becomes essential for a full description of the recognition process. The entropic effects of the solvent can be expected to be especially important for the interactions involving non-polar and aromatic amino acids [105]. The entropy of the solute can hardly be generalised, as the conformational entropy of both interacting protein and DNA molecules is heavily dependent on their composition and experimental conditions. Several studies have suggested entropy as the driving force of the specific sequence recognition in some complexes. Large increase in entropy was detected upon associacion of Arc repressor protein with its operator sequence, an effect which was not observed after non-specific interactions with other DNA molecules [106]. Entropic contributions are also essential for the binding of *Eco*R1 and *Bam*H1 restriction endonucleases to their specific DNA sequence targets [13]. Large "induced-fit" conformational changes taking place upon binding and interactions of intrinsically disordered proteins with DNA are exciting areas of active research, both working with energy landspaces heavily determined by the entropic effects [107–109].

de Ruiter and Zagrovic have recently explored the absolute binding free energies of the majority of pairs of amino acids with DNA and RNA bases by PMF calculations which include the entropic contributions [44]. They found that while stacked conformations with aromatic amino acids were the most favourable pairs in water environment, various hydrogen bond-featuring motifs provided the largest stabilisation energies when a lower dielectric constant (methanol) solvent was used. Notably, a guanine – aspartate pair adapting the conformation shown in Figure A.20a was found to be among the most favourable amino acid – DNA base combinations in the methanol environment. However, when the potentials of mean force were calculated in the water solvent, interactions with negatively charged amino acids were unanimously found unfavourable (displaying positive free energies of binding). Importantly, no significant amino acid interaction preferences towards any DNA base were found in this study. The binding free energies of most pairs in water were found to be around -2 kJ/mol. Interestingly, it was found that the binding free energies of interactions with guanine and cytosine are much more affected by the change of the dielectric constant of the environment than interactions with the other bases [44].

4.3 Comparison with protein – protein interactions

Berka *et al.* conducted an analysis of interaction energies in pairs of amino acids extracted from protein structures by methods analogous to those presented in this study. When analysing the pairs representative of the largest clusters found in each of the 20×20 distributions, it was found that, unless the amino acids carried the same charge, the respective cluster representatives provided universally stabilising interaction energies in the gas phase[2] [110]. On the other hand, there were multiple repulsive amino acid – DNA base interactions detected in my study, both by the force field as well as the CCSD(T)/CBS methods. It must be added, however, that my analysis was performed on all representatives, not only those found in largest clusters. Although the highest proportion of unfavourable contacts was found in the subset of pairs containing charged amino acids, some positive interaction energies were also found in contacts featuring polar or even aromatic amino acids. The origins of the repulsion may differ from contact to contact, although they are likely not primarily of electrostatic nature, as the isolated

[2]Some unfavourable interactions of the representative amino acids pairs were, however, detected when higher dielectric constant media were used [110].

DNA bases carry no charge.

Furthermore, interaction energy profiles of pairs of tryptophan with all other amino acids were prepared in their study. It was found that the representative contact of the largest cluster found in each distribution provided the most stabilising interaction energy for that pair [110]. No such behaviour was observed in my work on amino acid – DNA base contacts, in which the representative of the largest cluster (along with its other members) would provide interaction energies anywhere in the range of the interaction energy profile. It can be observed, however, that the interaction energy provided by the representative contact is often an average one observed among the members of its associated cluster (see, for example, Figures A.7d and A.8d).

Chapter 5

Conclusions

In this work I performed the most complete analysis of binding preferences in amino acid – DNA residue pairs to date. A large set of high-quality structures of protein – DNA complexes served as the basis for the extraction of the contacts. As many of the proteins in the set were homologous, bioinformatic tools were utilised to reduce the redundancy. Clustering of amino acid residues around the DNA bases in three dimensions was recognised in the distributions of the pairs and rigorously defined. This clustering was taken to be related to the functional role of individual amino acid – DNA nucleotide arrangements. These contacts were then subject to interaction energy calculations by a variety of methods.

Initially, the results of computationally less demanding MM methods utilising three commonly used FFs were compared with highly accurate *ab initio* benchmark interaction energies. These calculations were performed on a small set of representative amino acid – DNA base complexes. Very good correlations were found between the two sets of values for each FF, as long as highly strained complexes were excluded. The treatment of pairs in which the deformation energy was large, however, differed in the Amber and CHARMM class FFs.

Having investigated the computational reliability limits of the empirical methods, large-scale binding preferences of individual amino acid – DNA nucleotide pairs were probed. This was done by the investigation of the interaction energy profiles of the distributions and their comparison with the interaction energies provided by the members of the clusters found within. Specificity criteria which couple interaction energy characteristics of the pairs to their geometrical preferences, revealed by the clustering algorithm, were developed. These criteria potentially add new amino acid – DNA residue pairs to the library of specific contacts, expanding the usual definition of direct sequence recognition based on the complementary patterns of bidentate

hydrogen bonds. In addition to the canonical adenine – asparagine, adenine – glutamine and guanine – arginine pairs, several other combinations featuring charged, polar and aromatic amino acids were observed as being capable of uniquely distinguishing between individual DNA bases. Atomic-level description of each binding mode was provided, revealing that the respective direct sequence recognition can take place in any region around the DNA double helix. It must be added, however, that many of these results might be related to the gas phase approximation.

Finally, qualitative analysis of the electrostatic potential maps around individual DNA nucleotides moieties was used to explain the observed phenomenon of the phosphate group changing the non-covalent interaction properties the DNA bases. It was observed that an increase of electron density takes place around the DNA base moiety in the presence of the negatively charged group. Investigation of the electrostatic potentials of the previously identified sequence-specific amino acid – DNA residue complexes was performed afterwards, revealing that polarisation and charge transfer effects are an integral part of protein – DNA interactions.

Bibliography

[1] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond, "Crystal structure of the nucleosome resolution core particle at 2.8 Å resolution," *Nature*, vol. 389, no. 6648, pp. 251–260, 1997.

[2] S. Balasubramanian, F. Xu, and W. K. Olson, "DNA sequence-directed organization of chromatin: structure-based computational analysis of nucleosome-binding sequences," *Biophysical Journal*, vol. 96, no. 6, pp. 2245–2260, 2009.

[3] F. Battistini, C. A. Hunter, I. K. Moore, and J. Widom, "Structure-based identification of new high-affinity nucleosome binding sequences," *Journal of Molecular Biology*, vol. 420, no. 1-2, pp. 8–16, 2012.

[4] C. A. Davey, D. F. Sargent, K. Luger, A. W. Maeder, and T. J. Richmond, "Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution," *Journal of Molecular Biology*, vol. 319, no. 02, pp. 1097–1113, 2002.

[5] M. L. Smith, I.-T. Chen, Q. Zhan, P. M. O'Connor, and A. J. Fornace Jr, "Involvement of the p53 tumor suppressor in repair of uv-type DNA damage," *Oncogene*, vol. 10, no. 6, pp. 1053–1059, 1995.

[6] F. Drabløs, E. Feyzi, P. A. Aas, C. B. Vaagbø, B. Kavli, M. S. Bratlie, J. Peña-Diaz, M. Otterlei, G. Slupphaug, and H. E. Krokan, "Alkylation damage in DNA and RNA—repair mechanisms and medical significance," *DNA repair*, vol. 3, no. 11, pp. 1389–1407, 2004.

[7] L. Stojic, R. Brun, and J. Jiricny, "Mismatch repair and DNA damage signalling," *DNA repair*, vol. 3, no. 8, pp. 1091–1101, 2004.

[8] J. D. Watson and F. H. C. Crick, "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid," *Nature*, vol. 171, no. 4356, pp. 737–738, 1953.

[9] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, "The Protein Data Bank. A computer-based archival file for macromolecular structures," *European Journal of Biochemistry*, vol. 80, no. 2, pp. 319–324, 1977.

[10] G. Badis, E. T. Chan, H. van Bakel, L. Pena-Castillo, D. Tillo, K. Tsui, C. D. Carlson, A. J. Gossett, M. J. Hasinoff, C. L. Warren, M. Gebbia, S. Talukder, A. Yang, S. Mnaimneh, D. Terterov, D. Coburn, A. Li Yeo, Z. X. Yeo, N. D. Clarke, J. D. Lieb, A. Z. Ansari, C. Nislow, and T. R. Hughes, "A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters," *Molecular Cell*, vol. 32, no. 6, pp. 878–887, 2008.

[11] C. Zhu, K. J. R. P. Byers, R. P. McCord, Z. Shi, M. F. Berger, D. E. Newburger, K. Saulrieta, Z. Smith, M. V. Shah, M. Radhakrishnan, A. A. Philippakis, Y. Hu, F. De Masi, M. Pacek, A. Rolfs, T. Murthy, J. Labaer, and M. L. Bulyk, "High-resolution DNA-binding specificity analysis of yeast transcription factors," *Genome Research*, vol. 19, no. 4, pp. 556–566, 2009.

[12] C. J. Morton and J. E. Ladbury, "Water-mediated protein-DNA interactions: the relationship of thermodynamics to structural detail," *Protein Science*, vol. 5, no. 10, pp. 2115–2118, 1996.

[13] L. Jen-jacobson, L. E. Engler, J. T. Ames, M. R. Kurpiewski, and A. Grigorescu, "Thermodynamic Parameters of Specific and Nonspecific Protein-DNA Binding," *Supramolecular Chemistry*, vol. 12, pp. 143–160, 2000.

[14] R. Rohs, X. Jin, S. M. West, R. Joshi, B. Honig, and R. S. Mann, "Origins of specificity in protein - DNA recognition," *Annual Review of Biochemistry*, vol. 79, pp. 233–269, 2010.

[15] T. Gaj, C. A. Gersbach, and C. F. Barbas, "ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering," *Trends in Biotechnology*, vol. 31, no. 7, pp. 397–405, 2013.

[16] N. C. Seeman, J. M. Rosenberg, and A. Rich, "Sequence-specific recognition of double helical nucleic acids by proteins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 73, no. 3, pp. 804–808, 1976.

[17] A. V. Grishin, A. V. Alekseevsky, S. A. Spirin, and A. S. Kariagin, "Conserved structural features of ETS domain–DNA complexes," *Molekuliarnaia biologiia*, vol. 43, no. 4, pp. 666–674, 2009.

[18] Y. Kim, J. H. Geiger, S. Hahn, and P. B. Sigler, "Crystal structure of a yeast TBP/TATA-box complex," *Nature*, vol. 365, no. 6446, pp. 512–520, 1993.

[19] Z. Otwinowski, R. W. Schevitz, R. G. Zhang, C. L. Lawson, A. Joachimiak, R. Q. Marmorstein, B. F. Luisi, and P. B. Sigler, "Crystal structure of trp repressor/operator complex at atomic resolution," *Nature*, vol. 335, no. 6188, pp. 321–329, 1988.

[20] R. S. Hegde, S. R. Grossman, L. A. Laimins, and P. B. Sigler, "Crystal structure at 1.7 Å of the bovine papillomavirus-1 E2 DNA-binding domain bound to its DNA target," *Nature*, vol. 359, no. 6395, pp. 505–512, 1992.

[21] S. C. J. Parker, L. Hansen, H. O. Abaan, T. D. Tullius, and E. H. Margulies, "Local DNA topography correlates with functional noncoding regions of the human genome," *Science*, vol. 324, no. 5925, pp. 389–392, 2009.

[22] R. Rohs, S. M. West, P. Liu, and B. Honig, "Nuance in the double-helix and its role in protein - DNA recognition," *Current Opinion in Structural Biology*, vol. 19, no. 2, pp. 171–177, 2009.

[23] R. Rohs, S. M. West, A. Sosinsky, P. Liu, R. S. Mann, and B. Honig, "The role of DNA shape in protein - DNA recognition," *Nature*, vol. 461, no. 7268, pp. 1248–1253, 2009.

[24] Z. Shakked, G. Guerstein-Guzikevich, M. Eisenstein, F. Frolow, and D. Rabinovich, "The conformation of the DNA double helix in the crystal is dependent on its environment," *Nature*, vol. 342, no. 6248, pp. 456–460, 1989.

[25] S. Jones, P. van Heyningen, H. M. Berman, and J. M. Thornton, "Protein-DNA interactions: A structural analysis," *Journal of Molecular Biology*, vol. 287, no. 5, pp. 877–896, 1999.

[26] M. F. Berger, G. Badis, A. R. Gehrke, S. Talukder, A. A. Philippakis, L. Peña Castillo, T. M. Alleyne, S. Mnaimneh, O. B. Botvinnik, E. T. Chan, F. Khalid, W. Zhang, D. Newburger, S. a. Jaeger, Q. D. Morris, M. L. Bulyk, and T. R. Hughes, "Variation in Homeodomain

DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences," *Cell*, vol. 133, no. 7, pp. 1266–1276, 2008.

[27] P. Thomas and S. Podder, "Specificity in protein—nucleic acid interaction," *FEBS Letters*, vol. 96, no. 1, pp. 90–94, 1978.

[28] E. Akinrimisi and P. O. P. Ts'o, "Interactions of Purine with Proteins and Amino Acids," *Biochemistry*, vol. 3, no. 5, pp. 619–626, 1964.

[29] Y. Mandel-Gutfreund and H. Margalit, "Quantitative parameters for amino acid - base interaction: implications for prediction of protein - DNA binding sites," *Nucleic Acids Research*, vol. 26, no. 10, pp. 2306–2312, 1998.

[30] N. M. Luscombe, R. A. Laskowski, and J. M. Thornton, "Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level," *Nucleic Acids Research*, vol. 29, no. 13, pp. 2860–2874, 2001.

[31] M. M. Hoffman, M. A. Khrapov, C. J. Cox, J. Yao, L. Tong, and A. D. Ellington, "AANT: the Amino Acid - Nucleotide Interaction Database," *Nucleic Acids Research*, vol. 32, no. Database issue, pp. 174–181, 2004.

[32] T. Norambuena and F. Melo, "The Protein-DNA Interface database," *BMC Bioinformatics*, vol. 11, p. 262, 2010.

[33] B. Contreras-Moreira, "3D-footprint: a database for the structural analysis of protein-DNA complexes," *Nucleic Acids Research*, vol. 38, no. Database, pp. D91–D97, 2010.

[34] P. Prabakaran, J. An, M. M. Gromiha, S. Selvaraj, H. Uedaira, H. Kono, and a. Sarai, "Thermodynamic database for protein-nucleic acid interactions (ProNIT)," *Bioinformatics*, vol. 17, no. 11, pp. 1027–1034, 2001.

[35] S. Kiliç, E. R. White, D. M. Sagitova, J. P. Cornish, and I. Erill, "CollecTF: A database of experimentally validated transcription factor-binding sites in Bacteria," *Nucleic Acids Research*, vol. 42, no. Database, pp. 156–160, 2014.

[36] E. Wingender, P. Dietze, H. Karas, and R. Knüppel, "TRANSFAC: A database on transcription factors and their DNA binding sites," *Nucleic Acids Research*, vol. 24, no. 1, pp. 238–241, 1996.

[37] D. D. Kirsanov, O. N. Zanegina, E. A. Aksianov, S. A. Spirin, A. S. Karyagina, and A. V. Alexeevski, "NPIDB: Nucleic acid-Protein Interaction DataBase," *Nucleic Acids Research*, vol. 41, no. Database, pp. D517–D523, 2013.

[38] R. Bonaccorsi, A. Pullman, E. Scrocco, and J. Tomasi, "The molecular electrostatic potentials for the nucleic acid bases: Adenine, thymine, and cytosine," *Theoretica Chimica Acta*, vol. 24, no. 1, pp. 51–60, 1972.

[39] D. Perahia and A. Pullman, "The molecular electrostatic potentials of the complementary base pairs of DNA," *Theoretica Chimica Acta*, vol. 48, no. 3, pp. 263–266, 1978.

[40] J. Šponer and P. Hobza, "Nonplanar geometries of DNA bases. Ab initio second-order Moeller-Plesset study," *The Journal of Physical Chemistry*, vol. 98, no. 12, pp. 3161–3164, 1994.

[41] P. Hobza and J. Šponer, "Toward true DNA base-stacking energies: MP2, CCSD(T), and complete basis set calculations," *Journal of the American Chemical Society*, vol. 124, no. 39, pp. 11802–11808, 2002.

[42] J. Šponer, P. Jurečka, and P. Hobza, "Accurate interaction energies of hydrogen-bonded nucleic acid base pairs," *Journal of the American Chemical Society*, vol. 126, no. 32, pp. 10142–10151, 2004.

[43] P. Jurečka, J. Šponer, J. Černý, and P. Hobza, "Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs," *Physical Chemistry Chemical Physics: PCCP*, vol. 8, no. 17, pp. 1985–1993, 2006.

[44] A. de Ruiter and B. Zagrovic, "Absolute binding-free energies between standard RNA/DNA nucleobases and amino-acid sidechain analogs in different environments," *Nucleic Acids Research*, vol. 43, no. 2, pp. 708–718, 2014.

[45] F. Pichierri, M. Aida, M. M. Gromiha, and A. Sarai, "Free-Energy Maps of Base - Amino Acid Interactions for DNA - Protein Recognition," *Journal of the American Chemical Society*, vol. 121, no. 6, pp. 6152–6157, 1999.

[46] J. Singh and J. M. Thornton, *Atlas of Protein Side-Chain Interactions*, vol. I and II. Oxford: IRL Press, 1992.

[47] J. Singh and J. M. Thornton, "SIRIUS: an automated method for the analysis of the preferred packing arrangements between protein groups," *Journal of Molecular Biology*, vol. 211, no. 3, pp. 595–615, 1990.

[48] H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Buckhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J. D. Westbrook, and C. Zardecki, "The Protein Data Bank," *Acta Crystallographica D*, vol. 58, pp. 899–907, 2002.

[49] A. L. Morris, M. W. MacArthur, E. G. Hutchinson, and J. M. Thornton, "Stereochemical quality of protein structure coordinates," *Proteins: Structure, Function and Genetics*, vol. 12, no. 4, pp. 345–364, 1992.

[50] G. Wang and R. L. Dunbrack, "PISCES: A protein sequence culling server," *Bioinformatics*, vol. 19, no. 12, pp. 1589–1591, 2003.

[51] A. Bondi, "van der Waals Volumes and Radii," *The Journal of Physical Chemistry*, vol. 68, no. 3, pp. 441–451, 1964.

[52] W. Humphrey, A. Dalke, and K. Schulten, "VMD: Visual molecular dynamics," *Journal of Molecular Graphics*, vol. 14, pp. 33–38, 1996.

[53] D. Jakubec, J. Hostaš, R. A. Laskowski, P. Hobza, and J. Vondrášek, "Large-Scale Quantitative Assessment of Binding Preferences in Protein–Nucleic Acid Complexes," *Journal of Chemical Theory and Computation*, vol. 11, no. 4, pp. 1939–1948, 2015.

[54] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

[55] M. P. Styczynski, K. L. Jensen, I. Rigoutsos, and G. Stephanopoulos, "BLOSUM62 miscalculations improve search performance," *Nature Biotechnology*, vol. 26, no. 3, pp. 274–275, 2008.

[56] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, no. 22, pp. 10915–10919, 1992.

[57] P. Rice, I. Longden, and A. Bleasby, "EMBOSS: the European Molecular Biology Open Software Suite," *Trends in Genetics: TIG*, vol. 16, no. 6, pp. 276–277, 2000.

[58] E. W. Myers and W. Miller, "Optimal alignments in linear space," *Computer Applications in the Biosciences: CABIOS*, vol. 4, no. 1, pp. 11–17, 1988.

[59] J. Černý and P. Hobza, "Non-covalent interactions in biomacromolecules," *Physical Chemistry Chemical Physics: PCCP*, vol. 9, no. 39, pp. 5291–5303, 2007.

[60] J. Černý, M. Pitoňák, K. E. Riley, and P. Hobza, "Complete Basis Set Extrapolation and Hybrid Schemes for Geometry Gradients of Noncovalent Complexes," *Journal of Chemical Theory and Computation*, vol. 7, no. 12, pp. 3924–3934, 2011.

[61] K. Müller-Dethlefs and P. Hobza, "Noncovalent Interactions: A Challenge for Experiment and Theory," *Chemical Reviews*, vol. 100, no. 1, pp. 143–168, 2000.

[62] J. Šponer and P. Hobza, "MP2 and CCSD(T) study on hydrogen bonding, aromatic stacking and nonaromatic stacking," *Chemical Physics Letters*, vol. 267, no. 3-4, pp. 263–270, 1997.

[63] A. Halkier, T. Helgaker, P. Jø rgensen, W. Klopper, H. Koch, J. Olsen, and A. K. Wilson, "Basis-set convergence in correlated calculations on Ne, N2, and H2O," *Chemical Physics Letters*, vol. 286, no. 3-4, pp. 243–252, 1998.

[64] A. Halkier, T. Helgaker, P. Jø rgensen, W. Klopper, and J. Olsen, "Basis-set convergence of the energy in molecular Hartree–Fock calculations," *Chemical Physics Letters*, vol. 302, no. 5-6, pp. 437–446, 1999.

[65] O. Berger, O. Edholm, and F. Jähnig, "Molecular dynamics simulations of a fluid bilayer of dipalmitoylphosphatidylcholine at full hydration, constant pressure, and constant temperature," *Biophysical Journal*, vol. 72, no. May 1997, pp. 2002–2013, 1997.

[66] L. Celik, J. D. D. Lund, and B. Schiø tt, "Conformational dynamics of the estrogen receptor alpha: molecular dynamics simulations of the influence of binding site structure on protein dynamics," *Biochemistry*, vol. 46, pp. 1743–1758, 2007.

[67] A. Pérez, F. J. Luque, and M. Orozco, "Frontiers in molecular dynamics simulations of DNA," *Accounts of Chemical Research*, vol. 45, no. 2, pp. 196–205, 2012.

[68] A. Savelyev and G. A. Papoian, "Inter-DNA electrostatics from explicit solvent molecular dynamics simulations," *Journal of the American Chemical Society*, vol. 129, no. 19, pp. 6060–6061, 2007.

[69] A. D. Mackerell and L. Nilsson, "Molecular dynamics simulations of nucleic acid - protein complexes," *Current Opinion in Structural Biology*, vol. 18, no. 2, pp. 194–199, 2008.

[70] J. R. Maple, M.-J. Hwang, T. P. Stockfisch, U. Dinur, M. Waldman, C. S. Ewig, and A. T. Hagler, "Derivation of Class II Force Fields. I. Methodology and Quantum Force Field for the Alkyl Functional Group and Alkane Molecules," *Journal of Computational Chemistry*, vol. 15, no. 2, pp. 162–182, 1994.

[71] A. R. Leach, *Molecular modelling: principles and applications*. Pearson Education, 2001.

[72] J. W. Ponder and D. A. Case, "Force Fields for Protein Simulations," *Advances in Protein Chemistry*, vol. 66, pp. 27–86, 2003.

[73] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, "A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules," *Journal of the American Chemical Society*, vol. 117, no. 19, pp. 5179–5197, 1995.

[74] T. E. Cheatham, P. Cieplak, and P. A. Kollman, "A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat," *Journal of Biomolecular Structure & Dynamics*, vol. 16, no. 4, pp. 845–862, 1999.

[75] J. Wang, P. Cieplak, and P. A. Kollman, "How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?," *Journal of Computational Chemistry*, vol. 21, no. 12, pp. 1049–1074, 2000.

[76] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, and P. Kollman, "A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations," *Journal of Computational Chemistry*, vol. 24, no. 16, pp. 1999–2012, 2003.

[77] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus, "All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins," *The Journal of Physical Chemistry B*, vol. 102, no. 18, pp. 3586–3616, 1998.

[78] A. D. Mackerell, N. Banavali, and N. Foloppe, "Development and current status of the CHARMM force field for nucleic acids," *Biopolymers*, vol. 56, no. 4, pp. 257–265, 2001.

[79] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, "Comparison of multiple Amber force fields and development of improved protein backbone parameters," *Proteins*, vol. 65, no. 3, pp. 712–725, 2006.

[80] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, "Improved side-chain torsion potentials for the Amber ff99SB protein force field," *Proteins*, vol. 78, no. 8, pp. 1950–1958, 2010.

[81] N. Foloppe and A. D. J. MacKerell, "All-Atom Empirical Force Field for Nucleic Acids : I . Parameter Optimization Based on Small Molecule and Condensed Phase Macromolecular Target Data," *Journal of Computational Chemistry*, vol. 21, no. 2, pp. 86–104, 2000.

[82] P. Cieplak, F.-Y. Dupradeau, Y. Duan, and J. Wang, "Polarization effects in molecular mechanical force fields," *Journal of Physics: Condensed Matter*, vol. 21, no. 33, p. 333102, 2009.

[83] R. S. Mulliken, "Electronic Population Analysis on LCAO-MO Molecular Wave Functions. I," *The Journal of Chemical Physics*, vol. 23, no. 10, p. 1833, 1955.

[84] K. Berka, R. A. Laskowski, K. E. Riley, P. Hobza, and J. Vondrášek, "Representative amino acid side chain interactions in proteins. A comparison of highly accurate correlated ab initio quantum chemical and empirical potential procedures," *Journal of Chemical Theory and Computation*, vol. 5, no. 4, pp. 982–992, 2009.

[85] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, "UCSF Chimera - A visualization system for exploratory research and analysis," *Journal of Computational Chemistry*, vol. 25, pp. 1605–1612, 2004.

[86] B. Hess, C. Kutzner, D. Van Der Spoel, and E. Lindahl, "GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation," *Journal of Chemical Theory and Computation*, vol. 4, pp. 435–447, 2008.

[87] J. S. Murray, Z. Peralta-Inga, P. Politzer, K. Ekanayake, and P. Lebreton, "Computational characterization of nucleotide bases: molecular surface electrostatic potentials and local ionization energies, and local polarization energies," *International Journal of Quantum Chemistry*, vol. 83, no. 1, pp. 245–254, 2001.

[88] F. B. Sheinerman and B. Honig, "On the role of electrostatic interactions in the design of protein-protein interfaces," *Journal of Molecular Biology*, vol. 318, no. 1, pp. 161–177, 2002.

[89] F. B. Sheinerman, R. Norel, and B. Honig, "Electrostatic aspects of protein-protein interactions," *Current Opinion in Structural Biology*, vol. 10, no. 2, pp. 153–159, 2000.

[90] I. T. Weber and T. A. Steitz, "Model of specific complex between catabolite gene activator protein and B-DNA suggested by electrostatic complementarity," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 81, no. 13, pp. 3973–3977, 1984.

[91] S. Jones, H. P. Shanahan, H. M. Berman, and J. M. Thornton, "Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins," *Nucleic Acids Research*, vol. 31, no. 24, pp. 7189–7198, 2003.

[92] Y. Tsuchiya, K. Kinoshita, and H. Nakamura, "Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces," *Proteins: Structure, Function and Genetics*, vol. 55, no. 4, pp. 885–894, 2004.

[93] P. K. Weiner, R. Langridge, J. M. Blaney, R. Schaefer, and P. A. Kollman, "Electrostatic potential molecular surfaces," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 79, no. 12, pp. 3754–3758, 1982.

[94] M. L. Kopka, C. Yoon, D. Goodsell, P. Pjura, and R. E. Dickerson, "The molecular origin of DNA-drug specificity in netropsin and distamycin," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 82, no. 5, pp. 1376–1380, 1985.

[95] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox, "Gaussian 09 Revision D.01." Gaussian Inc. Wallingford CT 2009.

[96] R. A. Kendall, T. H. Jr., Dunnin, and R. J. Harrison, "Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions," *Journal of Chemical Physics*, vol. 96, no. 9, p. 6796, 1992.

[97] G. Schaftenaar and J. H. Noordik, "Molden: a pre- and post-processing program for molecular and electronic structures," *Journal of Computer-Aided Molecular Design*, vol. 14, no. 2, pp. 123–134, 2000.

[98] A. Pérez, I. Marchán, D. Svozil, J. Sponer, T. E. Cheatham, C. A. Laughton, and M. Orozco, "Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers," *Biophysical Journal*, vol. 92, no. 11, pp. 3817–3829, 2007.

[99] D. Freedman and P. Diaconis, "On the histogram as a density estimator: L2 theory," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 57, no. 4, pp. 453–476, 1981.

[100] V. Tsui and D. A. Case, "Molecular dynamics simulations of nucleic acids with a generalized born solvation model," *Journal of the American Chemical Society*, vol. 122, no. 11, pp. 2489–2498, 2000.

[101] T. Gaillard and D. A. Case, "Evaluation of DNA force fields in implicit solvation," *Journal of Chemical Theory and Computation*, vol. 7, no. 10, pp. 3181–3198, 2011.

[102] J. Chocholoušová and M. Feig, "Implicit solvent simulations of DNA and DNA-protein complexes: Agreement with explicit solvent vs experiment," *Journal of Physical Chemistry B*, vol. 110, no. 34, pp. 17240–17251, 2006.

[103] F. Fogolari, A. Corazza, and G. Esposito, "Accuracy assessment of the linear Poisson-Boltzmann equation and reparametrization of the OBC generalized Born model for nucleic acids and nucleic acid-protein complexes," *Journal of Computational Chemistry*, vol. 36, no. 9, pp. 585–596, 2015.

[104] P. V. Benos, M. L. Bulyk, and G. D. Stormo, "Additivity in protein - DNA interactions: how good an approximation is it?," *Nucleic Acids Research*, vol. 30, no. 20, pp. 4442–4451, 2002.

[105] J. M. Sturtevant, "Heat capacity and entropy changes in processes involving proteins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 6, pp. 2236–2240, 1977.

[106] D. Foguel and J. L. Silva, "Cold denaturation of a repressor-operator complex: the role of entropy in protein - DNA recognition," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 91, no. 17, pp. 8244–8247, 1994.

[107] R. S. Spolar and M. T. Record, "Coupling of local folding to site-specific binding of proteins to DNA," *Science*, vol. 263, no. 5148, pp. 777–784, 1994.

[108] M. Fuxreiter, I. Simon, and S. Bondos, "Dynamic protein-DNA recognition: Beyond what can be seen," *Trends in Biochemical Sciences*, vol. 36, no. 8, pp. 415–423, 2011.

[109] A. N. Naganathan and M. Orozco, "The conformational landscape of an intrinsically disordered DNA-binding domain of a transcription regulator," *Journal of Physical Chemistry B*, vol. 117, no. iii, pp. 13842–13850, 2013.

[110] K. Berka, R. A. Laskowski, P. Hobza, and J. Vondrášek, "Energy matrix of structurally important side-chain/side-chain interactions in proteins," *Journal of Chemical Theory and Computation*, vol. 6, no. 7, pp. 2191–2203, 2010.

Appendix A

Illustrations

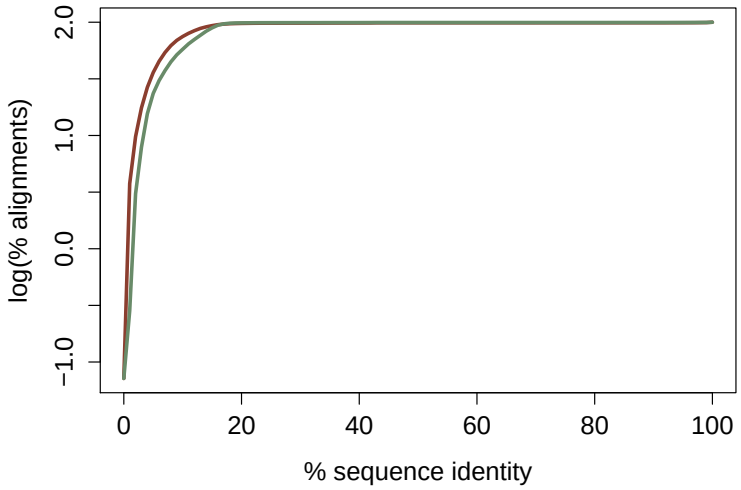


Figure A.1: Logarithm of the percent of alignments (y-axis) with sequence identity score at most $X\%$ (x-axis). Brown - *needle* tool, green - *stretcher* tool.

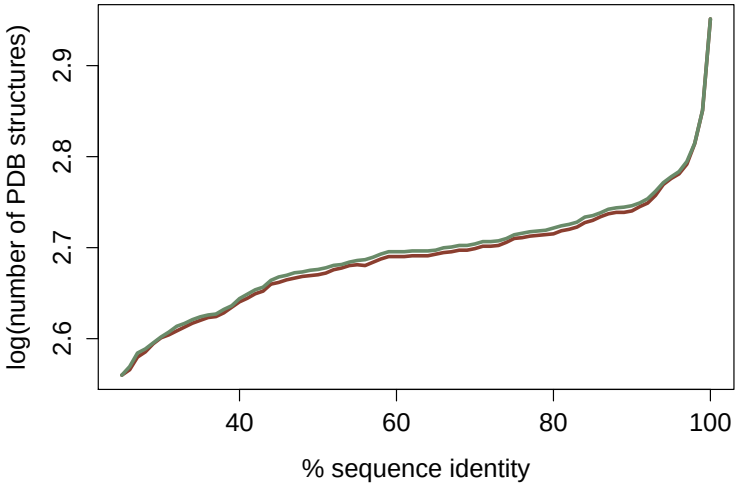
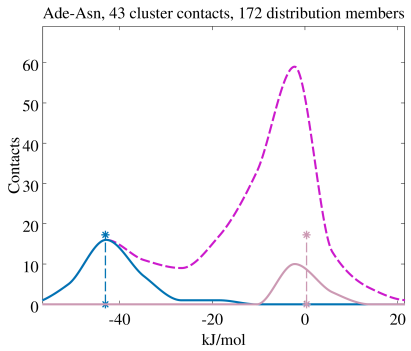
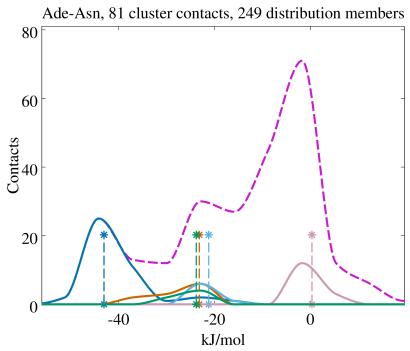


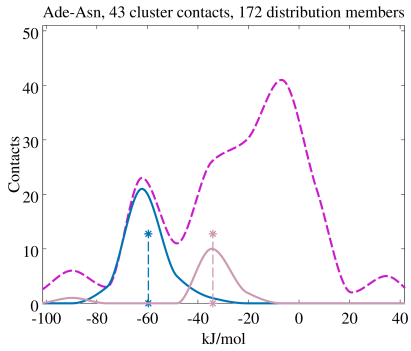
Figure A.2: The number of PDB structures left in the data set (shown as a logarithm, y-axis) as a function of maximal allowed % sequence identity of any pair of sequences (x-axis). Brown - "soft" approach, green - "hard" approach.



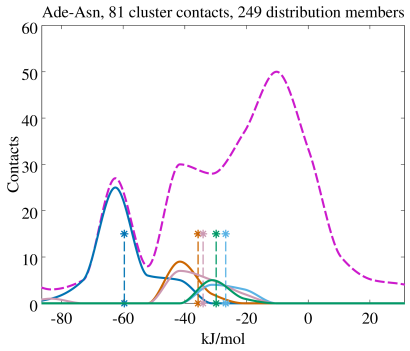
(a) adenine – asparagine, 90%



(b) adenine – asparagine, 100%

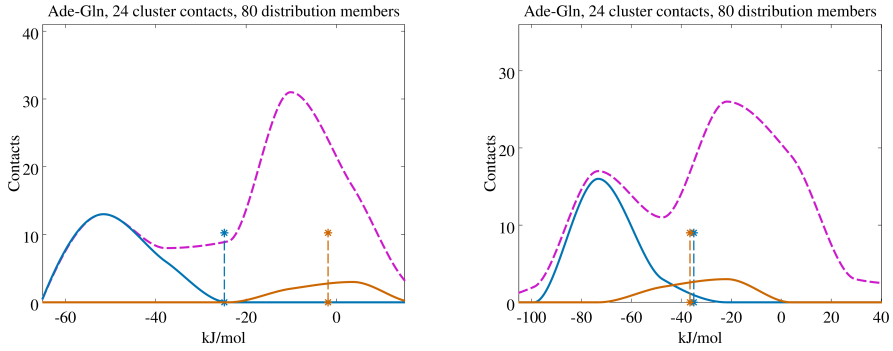


(c) dAMP – asparagine, 90%



(d) dAMP – asparagine, 100%

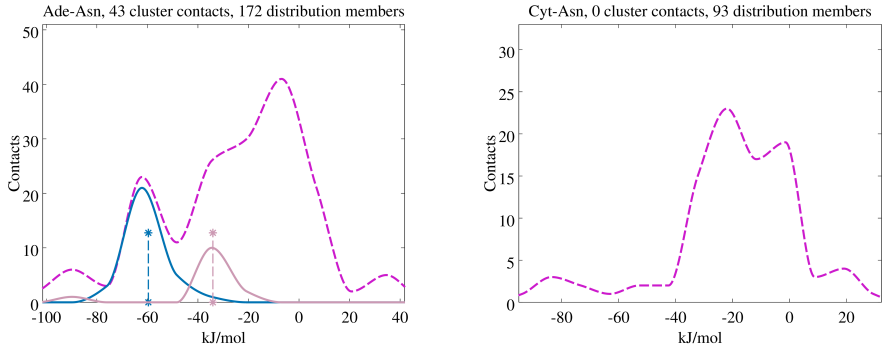
Figure A.3: Adenine (dAMP) – asparagine interaction energy profiles constructed at 90% and 100% maximum sequence identity levels. Dashed pink line - interaction energy profile created by considering all contacts in the distributions; solid profiles - energies of the cluster members; dashed vertical lines - energies of the respective cluster representatives.



(a) adenine – glutamine, 90%

(b) dAMP – glutamine, 90%

Figure A.4: Interaction energy profiles of adenine (dAMP) – glutamine pairs constructed at 90% sequence identity level. Only contacts featuring a direct interaction with the base moiety were considered. Color coding as in Figure A.3.



(a) dAMP – asparagine, 90%

(b) dCMP – asparagine, 90%

(c) dGMP – asparagine, 90%

(d) TMP – asparagine, 90%

Figure A.5: Interaction energy profiles of asparagine – dAMP, dCMP, dGMP and TMP pairs constructed at 90% sequence identity level. Only the complexes in which asparagine interacts directly with the base moiety were considered. Color coding as in Figure A.3.

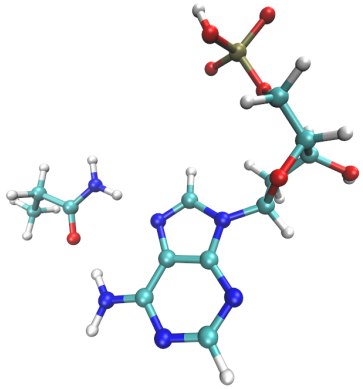
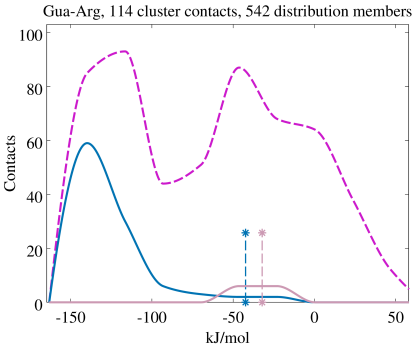
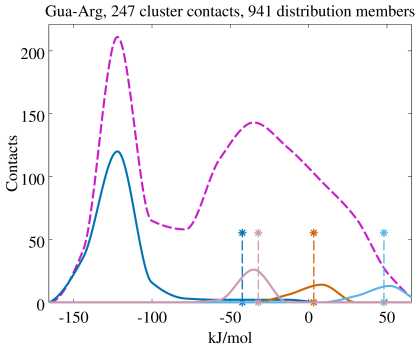


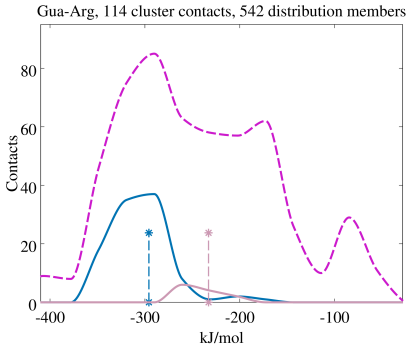
Figure A.6: The dAMP – asparagine dimer representative of the distinct low-lying cluster. Specific contacts with glutamine adapt the same geometry.



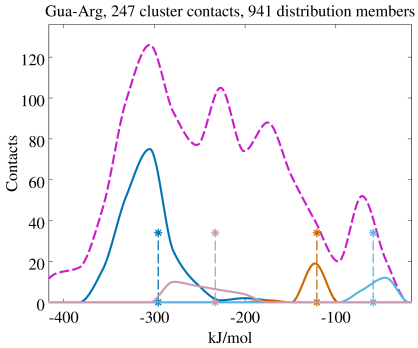
(a) guanine – arginine, 90%



(b) guanine – arginine, 100%

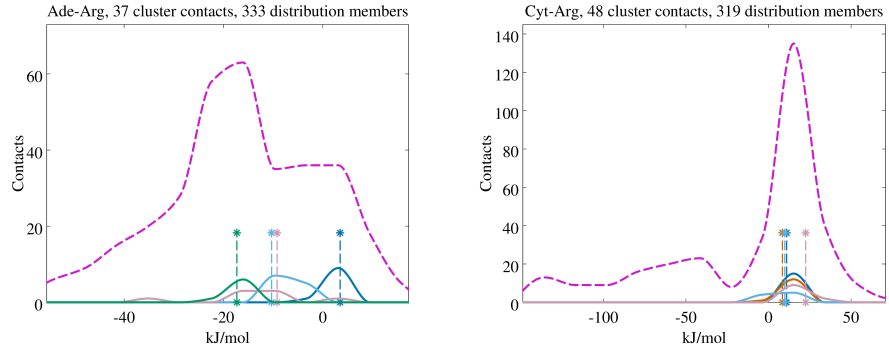


(c) dGMP – arginine, 90%



(d) dGMP – arginine, 100%

Figure A.7: Guanine (dGMP) – arginine interaction energy profiles constructed at 90% and 100% maximum sequence identity levels. Color coding as in Figure A.3.



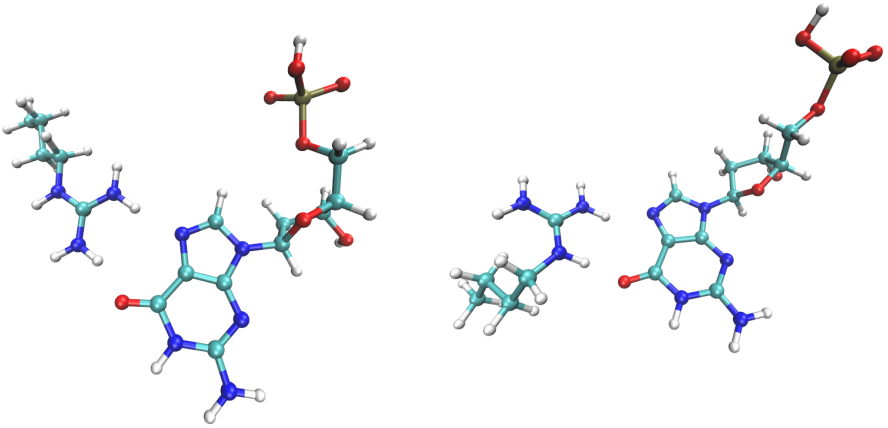
(a) adenine – arginine, 90%

(b) cytosine – arginine, 90%

(c) guanine – arginine, 90%

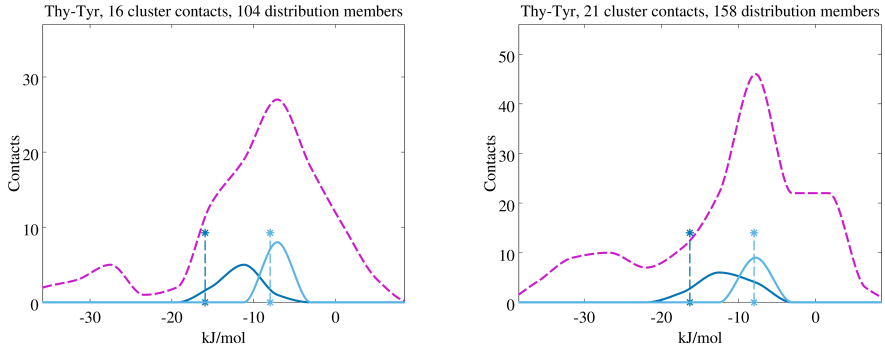
(d) thymine – arginine, 90%

Figure A.8: Interaction energy profiles of arginine – adenine, cytosine, guanine and thymine pairs constructed at 90% sequence identity level. Only the complexes in which arginine interacts directly with the base were considered. Color coding as in Figure A.3.



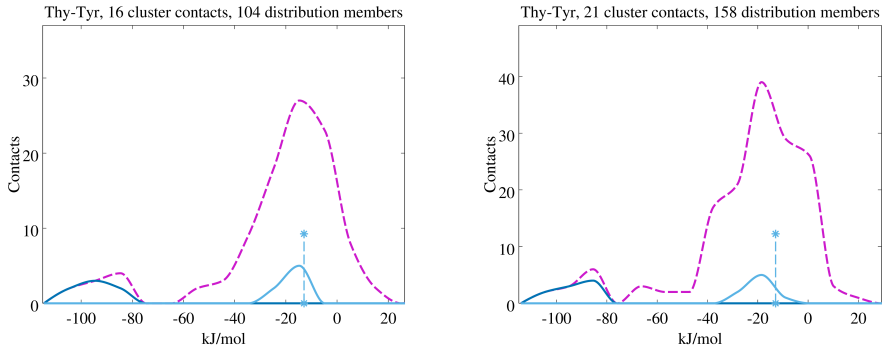
(a) Cluster conformation. (b) Alternative geometry.

Figure A.9: dGMP – arginine dimer representative of the distinct low-lying cluster and an alternative isoenergetic non-cluster geometry. Two other conformations are possible.



(a) thymine – tyrosine, 90%

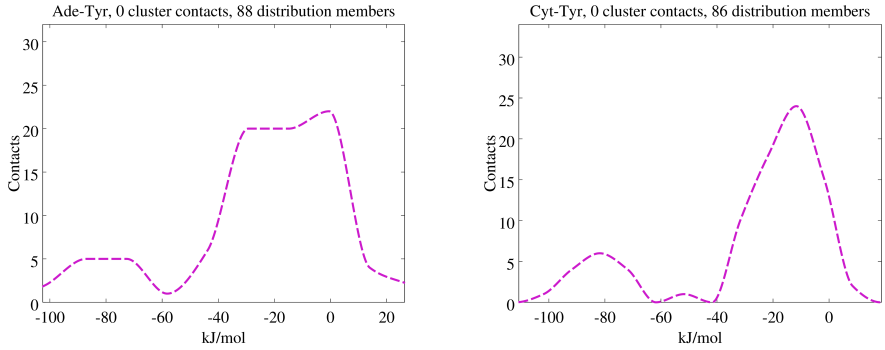
(b) thymine – tyrosine, 100%



(c) TMP – tyrosine, 90%

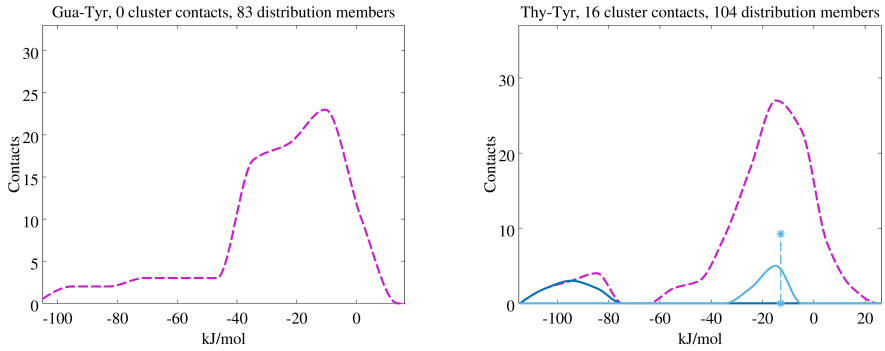
(d) TMP – tyrosine, 100%

Figure A.10: Thymine (TMP) – tyrosine interaction energy profiles constructed at 90% and 100% maximum sequence identity levels. Only the contacts involving interactions with the DNA base were included. Color coding as in Figure A.3.



(a) dAMP – tyrosine, 90%

(b) dCMP – tyrosine, 90%



(c) dGMP – tyrosine, 90%

(d) TMP – tyrosine, 90%

Figure A.11: Interaction energy profiles of tyrosine – dAMP, dCMP, dGMP and TMP pairs constructed at 90% sequence identity level. Only the complexes in which tyrosine interacts partially with the base were considered. Color coding as in Figure A.3.

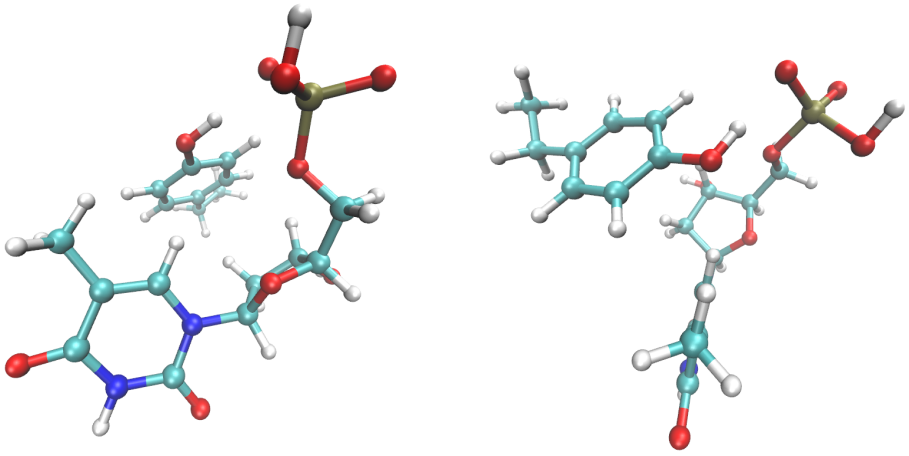


Figure A.12: TMP – tyrosine dimer chosen from the distinct low-lying cluster. Two views are presented.

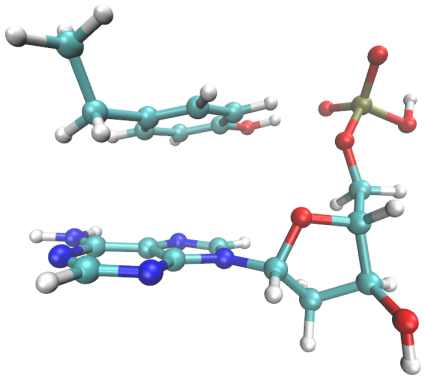


Figure A.13: dAMP – tyrosine dimer providing the same interaction energy as the pair shown in Figure A.12.

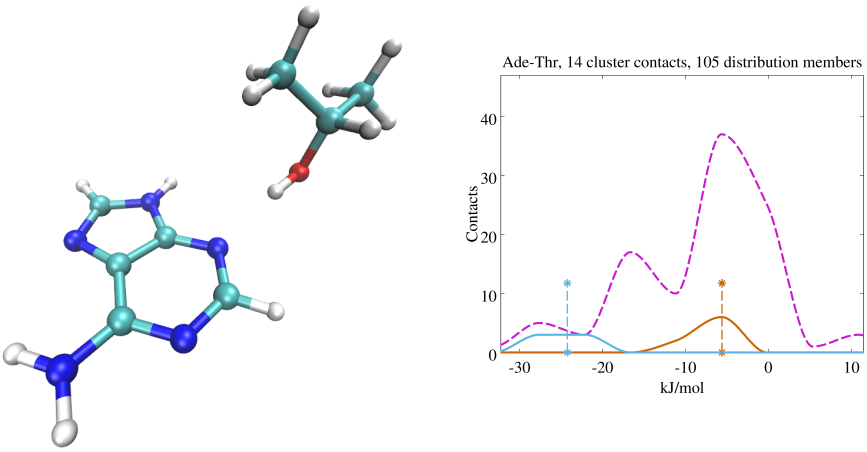


Figure A.14: Representative of the vaguely defined adenine – threonine cluster and the corresponding interaction energy profile constructed from contacts provided by non-identical protein chains (100% sequence identity criterium).

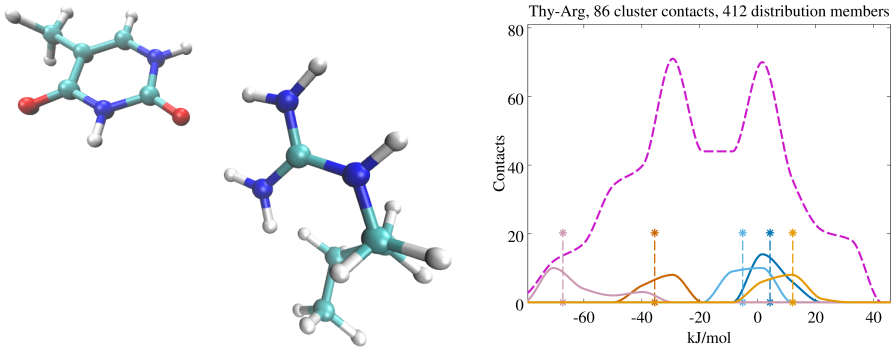


Figure A.15: Representative of the bluntly defined thymine – arginine cluster and the corresponding interaction energy profile constructed at 90% maximum allowed sequence identity.

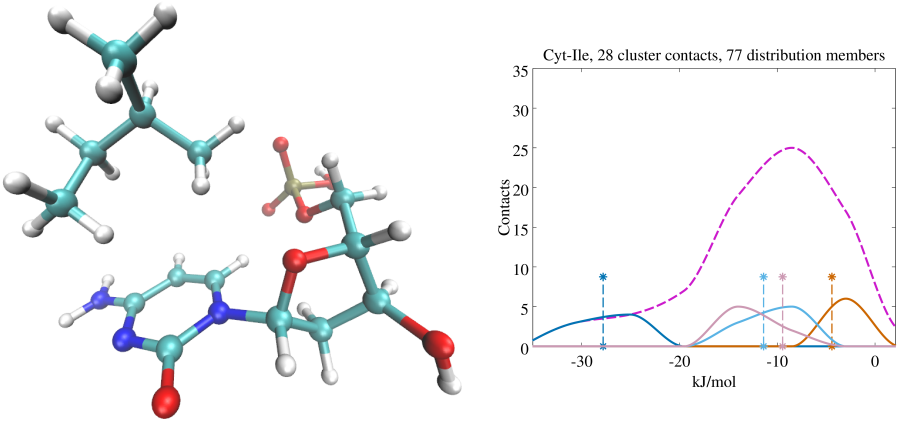
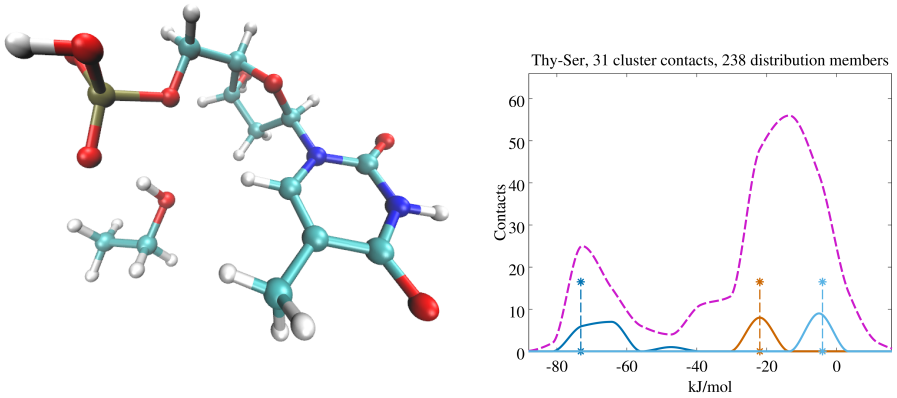
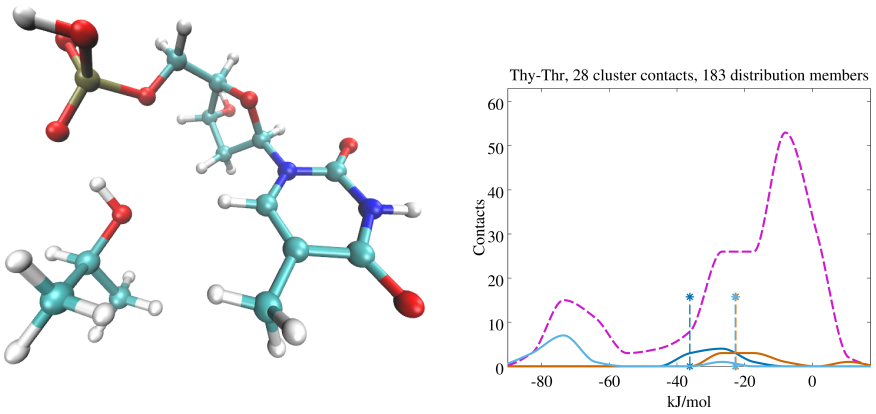


Figure A.16: Representative of the vaguely defined dCMP – isoleucine cluster and the corresponding interaction energy profile constructed from contacts provided by non-identical protein chains (100% sequence identity criterium).

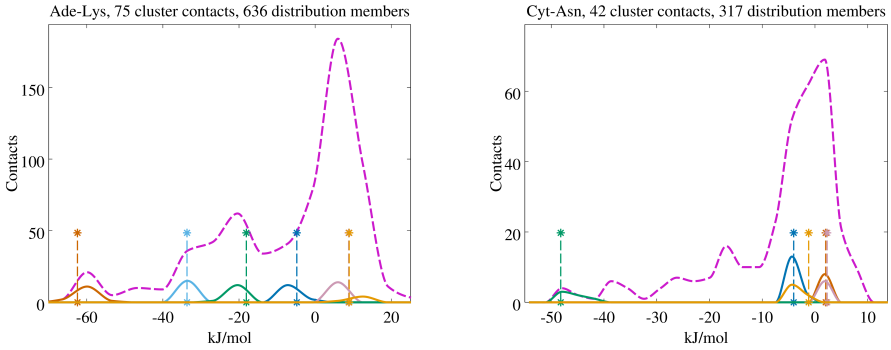


(a) TMP – serine



(b) TMP – threonine

Figure A.17: Representatives of the TMP – serine and threonine clusters and the corresponding interaction energy profiles constructed from contacts provided by non-identical protein chains (100% sequence identity criterium).



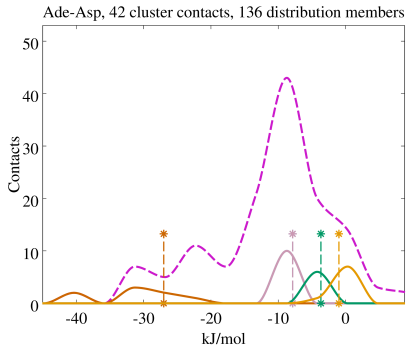
(a) adenine – lysine, 100%

(b) cytosine – asparagine, 100%

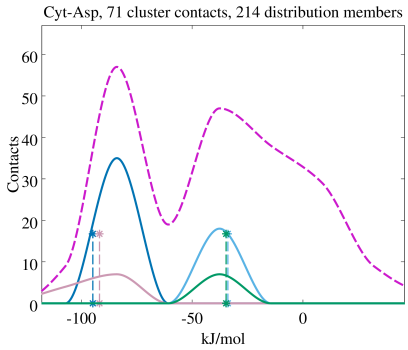
(c) cytosine – lysine, 100%

(d) guanine – aspartate, 100%

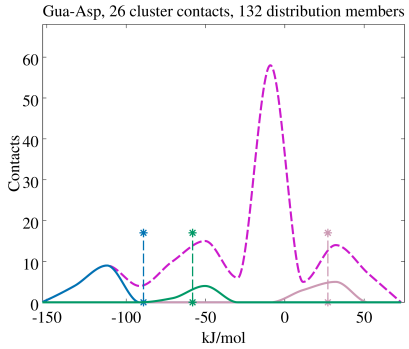
Figure A.18: Interaction energy profiles of adenine – lysine, cytosine – asparagine, cytosine – lysine and guanine – aspartate pairs constructed at 100% sequence identity level. All contacts were considered in the distributions. Color coding as in Figure A.3.



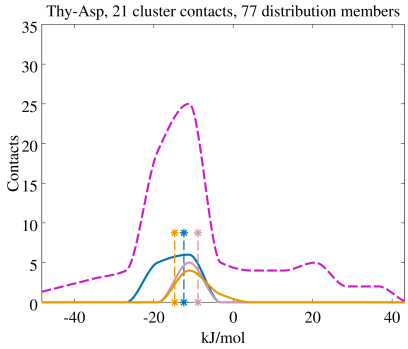
(a) Adenine – aspartate, 100%



(b) Cytosine – aspartate, 100%

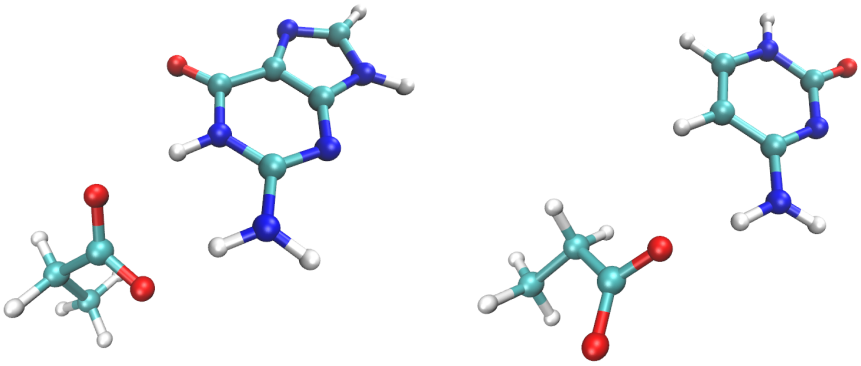


(c) Guanine – aspartate, 100%



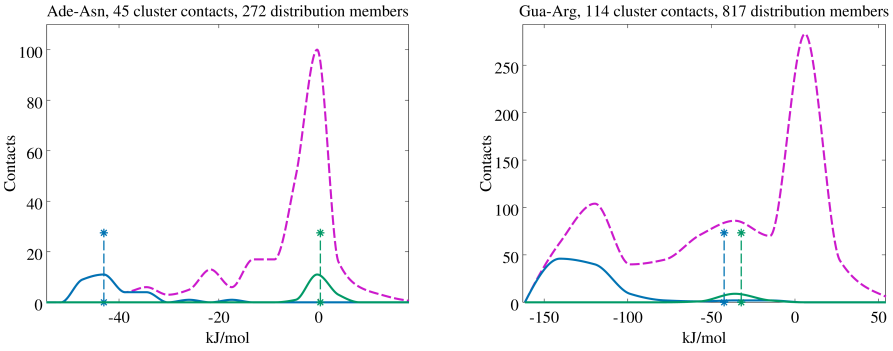
(d) Thymine – aspartate, 100%

Figure A.19: Interaction energy profiles of aspartate – adenine, cytosine, guanine and thymine pairs constructed after identical protein structures were discarded. Color coding as in Figure A.3.



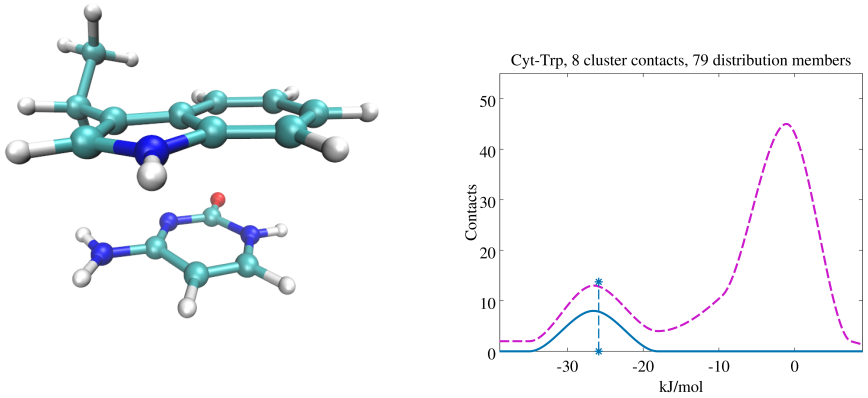
(a) Guanine – aspartate (b) Cytosine – aspartate

Figure A.20: Guanine – aspartate and cytosine – aspartate dimers chosen from the distinct low-lying clusters found in Figures A.19c and A.19b, respectively.

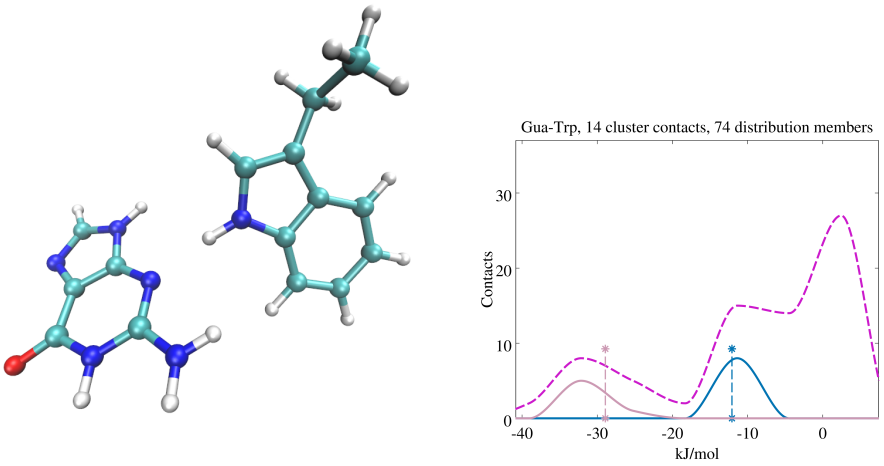


(a) adenine – asparagine, 90% (b) guanine – arginine, 90%

Figure A.21: Interaction energy profiles of adenine – asparagine and guanine – arginine pairs constructed at 90% sequence identity level. All contacts were considered in the distributions. Color coding as in Figure A.3.

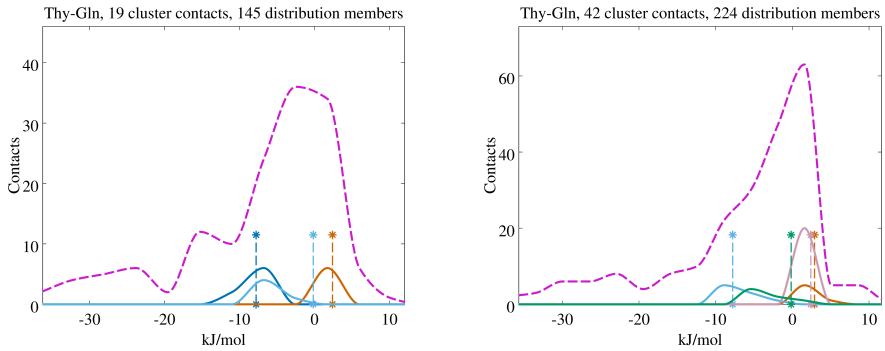


(a) cytosine – tryptophan



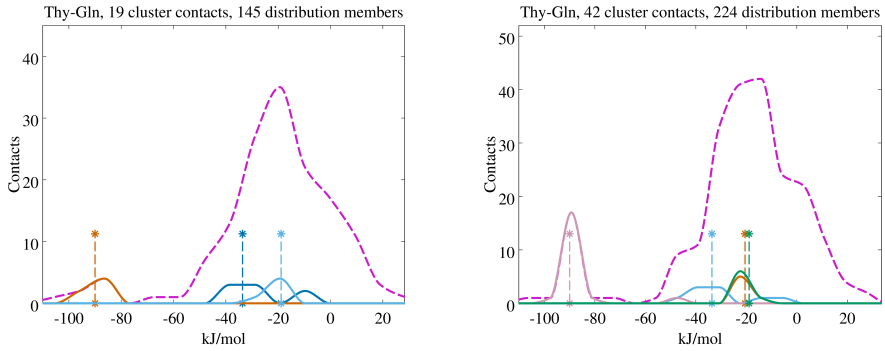
(b) guanine – tryptophan

Figure A.22: Representatives of the cytosine – tryptophan and guanine – tryptophan clusters and the corresponding interaction energy profiles constructed from contacts provided by non-identical protein chains (100% sequence identity criterium).



(a) thymine – glutamine, 90%

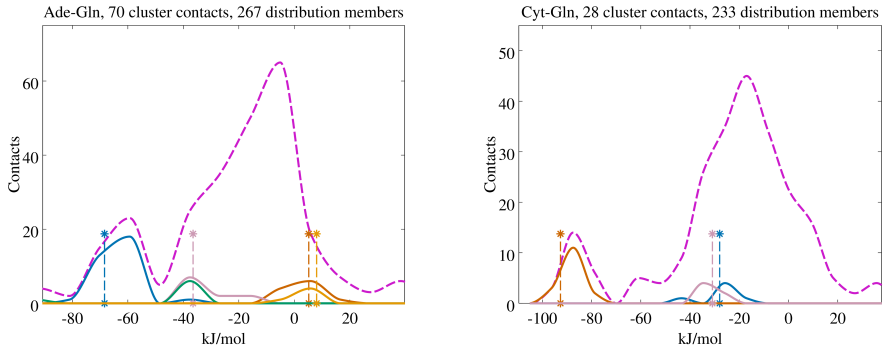
(b) thymine – glutamine, 100%



(c) TMP – glutamine, 90%

(d) TMP – glutamine, 100%

Figure A.23: Thymine (TMP) – glutamine interaction energy profiles constructed at 90% and 100% maximum sequence identity levels. Color coding as in Figure A.3. Same clusters share their coloration horizontally, but not vertically: orange in Figure A.23a corresponds to orange in Figure A.23c, but to the orange clusters in the second column.



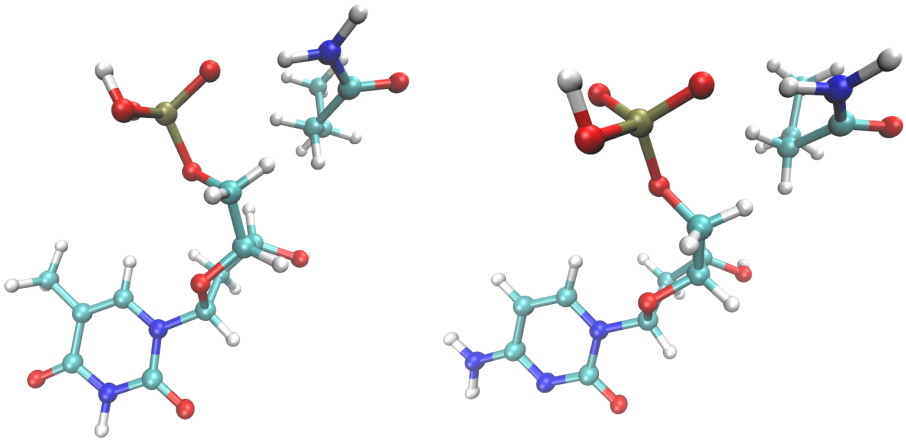
(a) dAMP – glutamine, 100%

(b) dCMP – glutamine, 100%

(c) dGMP – glutamine, 100%

(d) TMP – glutamine, 100%

Figure A.24: Interaction energy profiles of glutamine – dAMP, dCMP, dGMP and TMP pairs constructed after discarding structures of identical proteins. Color coding as in Figure A.3.



(a) TMP – glutamine (b) dCMP – glutamine

Figure A.25: TMP – glutamine and dCMP – glutamine dimers representative of the distinct low-lying clusters.

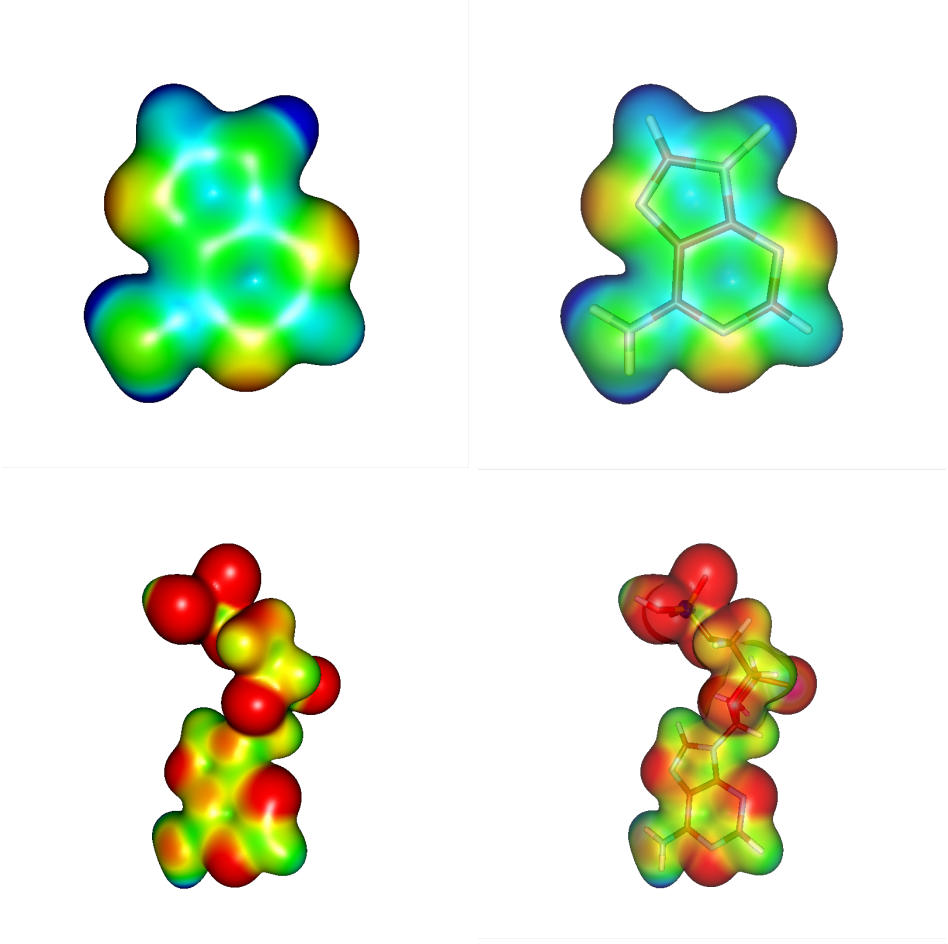


Figure A.26: Electrostatic potentials around adenine base and nucleotide forms. View from the top of the base. Contour value 0.01; color scale (in Volts): red < -0.10, yellow -0.05, green 0.00, light blue 0.05, blue > 0.10.

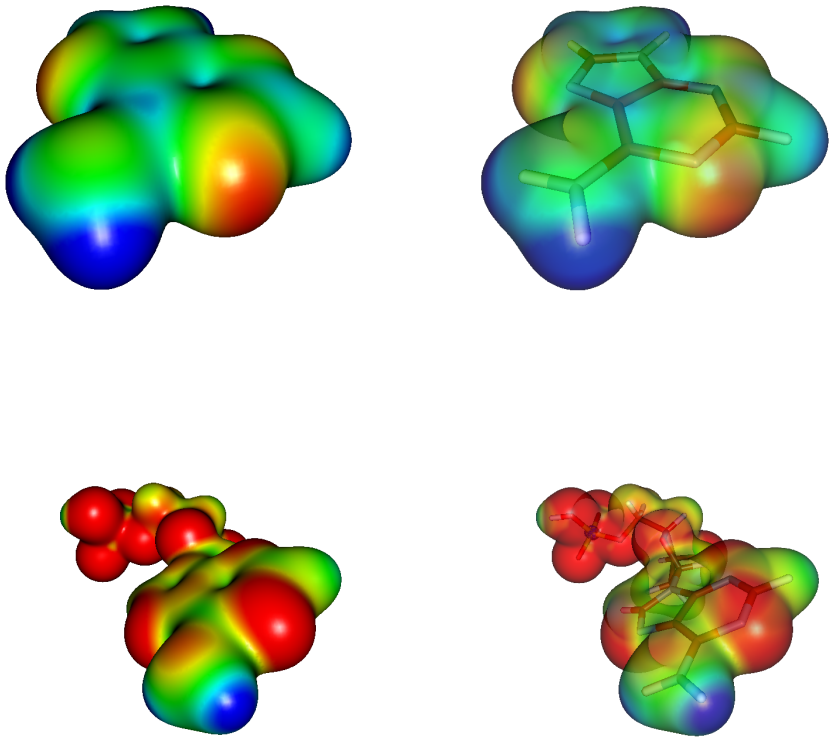


Figure A.27: Electrostatic potentials around adenine base and nucleotide forms. View from the Watson-Crick edge. Color coding as in Figure A.26.

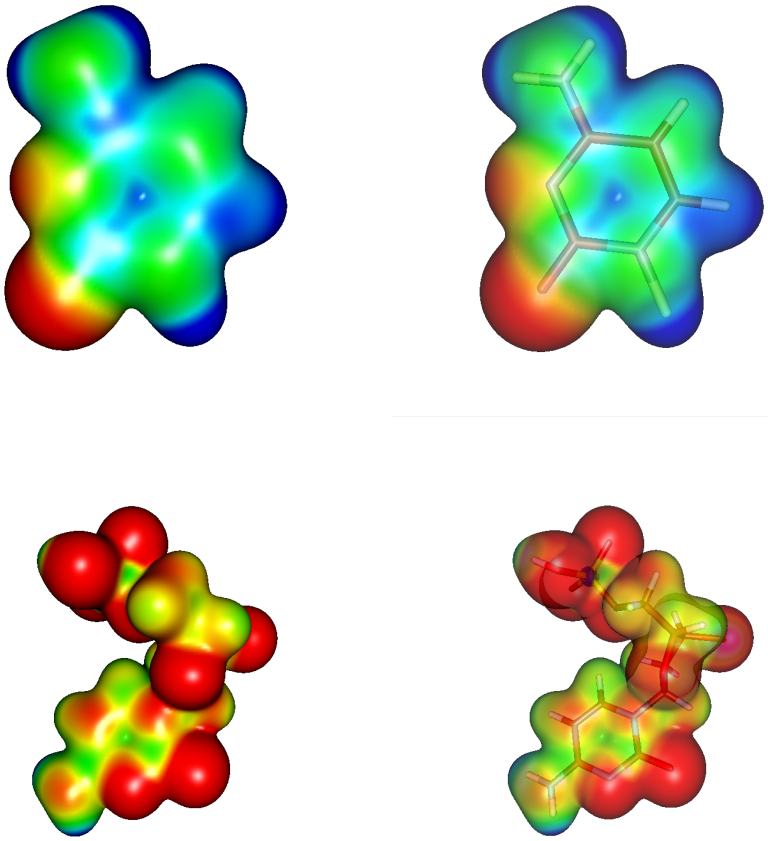


Figure A.28: Electrostatic potentials around cytosine base and nucleotide forms. View from the top of the base. Color coding as in Figure A.26.

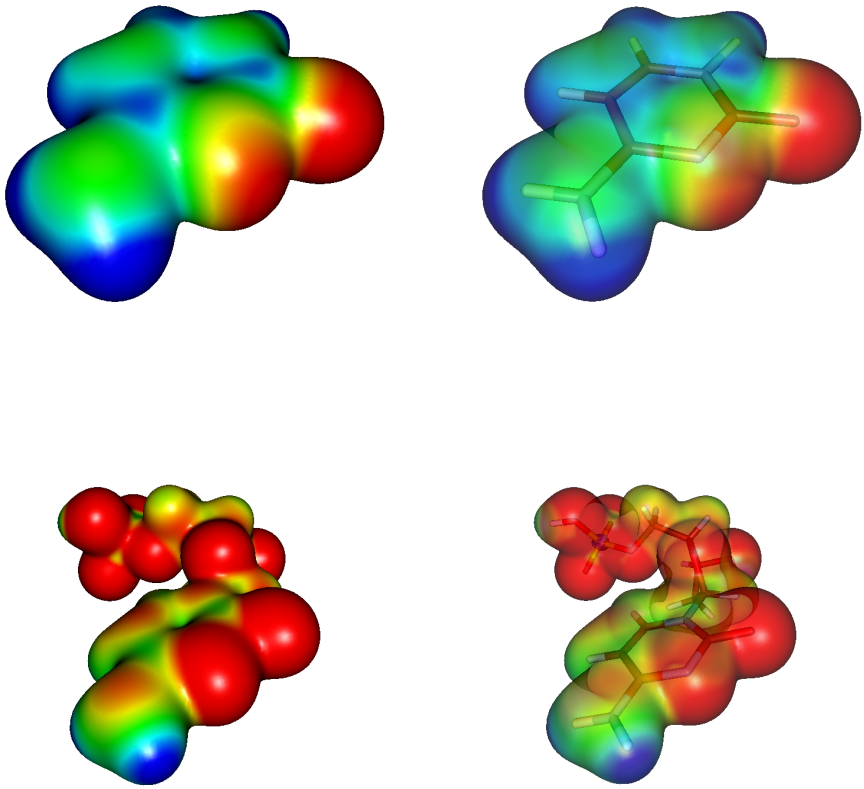


Figure A.29: Electrostatic potentials around cytosine base and nucleotide forms. View from the Watson-Crick edge. Color coding as in Figure A.26.

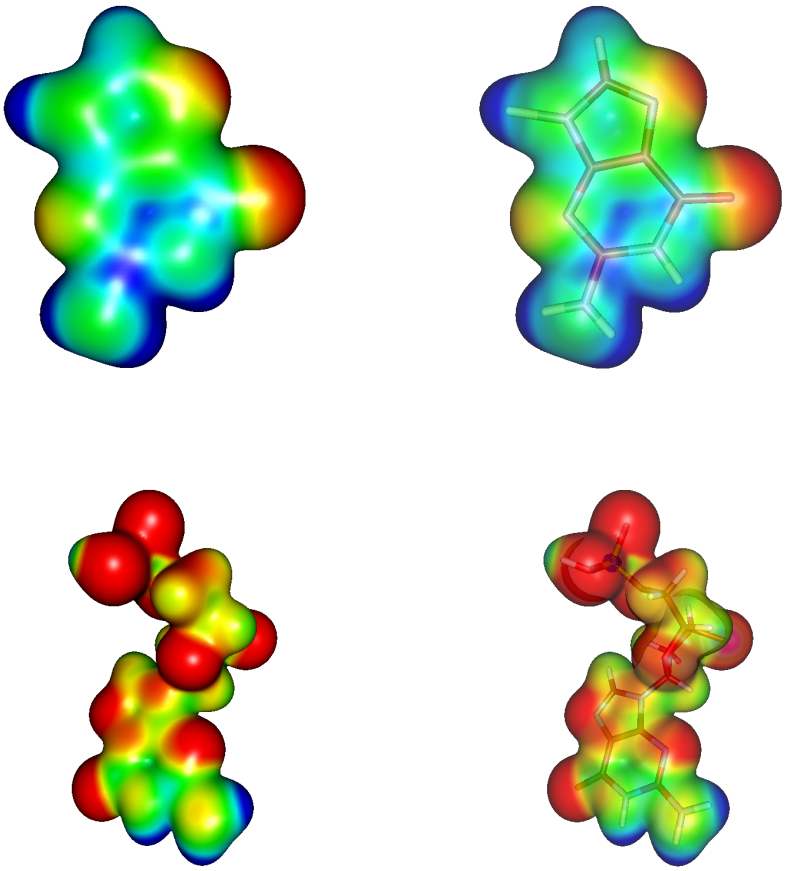


Figure A.30: Electrostatic potentials around guanine base and nucleotide forms. View from the top of the base. Color coding as in Figure A.26.

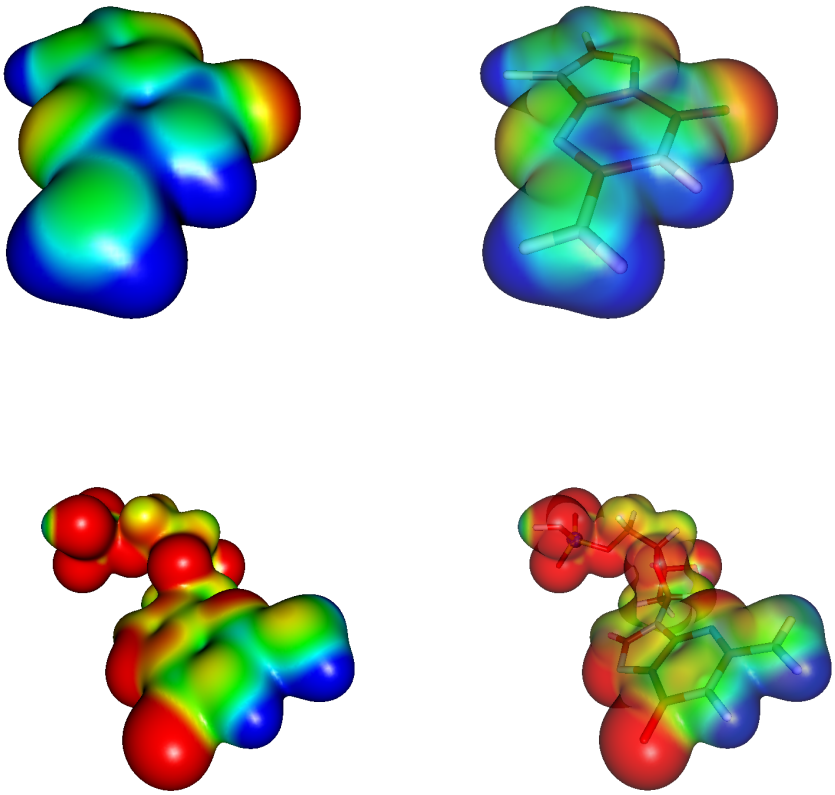


Figure A.31: Electrostatic potentials around guanine base and nucleotide forms. Note the different orientation of the base. View from the Watson-Crick edge. Color coding as in Figure A.26.

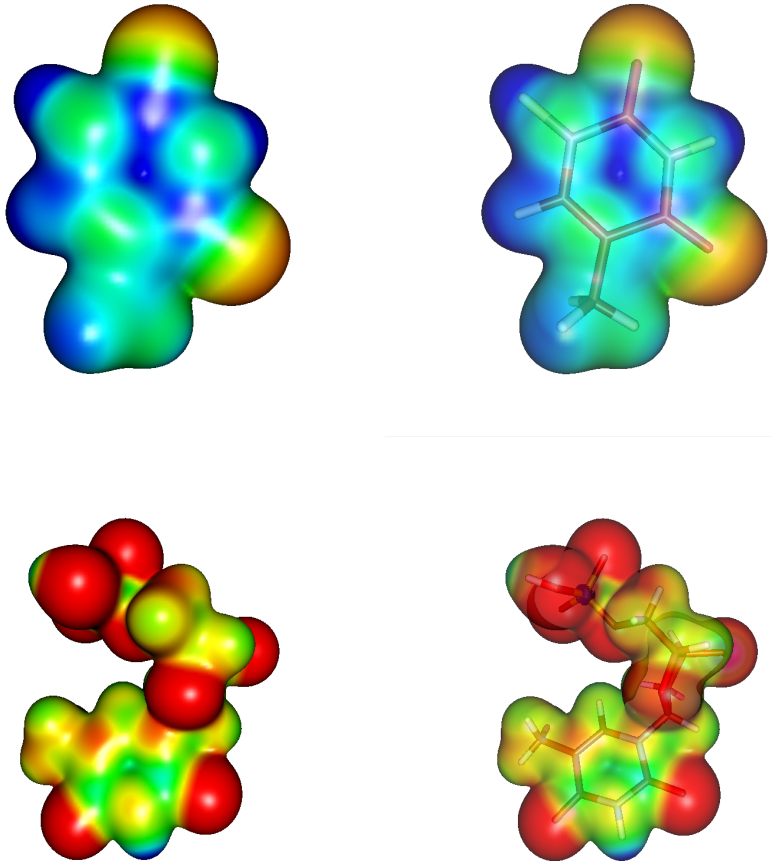


Figure A.32: Electrostatic potentials around thymine base and nucleotide forms. View from the top of the base. Color coding as in Figure A.26.

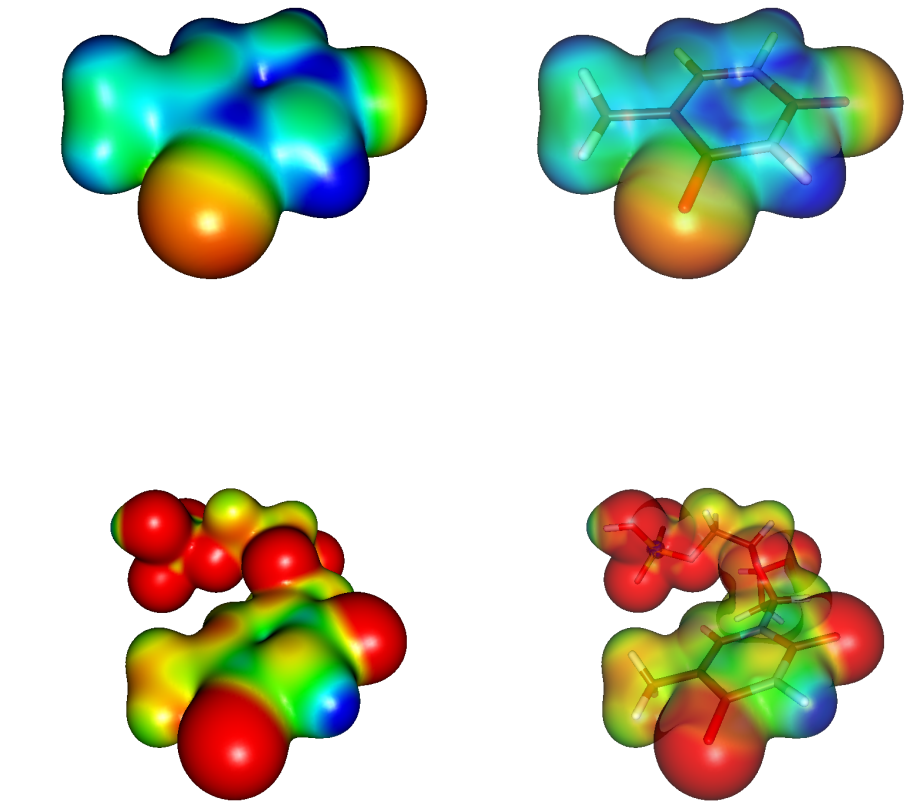


Figure A.33: Electrostatic potentials around thymine base and nucleotide forms. View from the Watson-Crick edge. Color coding as in Figure A.26.

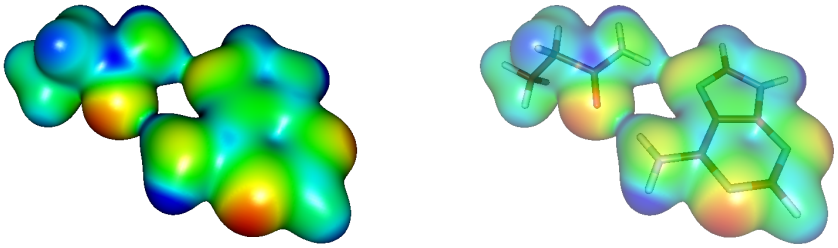


Figure A.34: Electrostatic potential of the sequence-specific adenine – asparagine pair. Color coding as in Figure A.26.

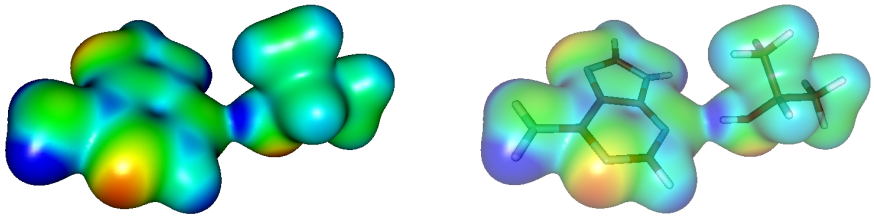


Figure A.35: Electrostatic potential of the sequence-specific adenine – threonine pair. Color coding as in Figure A.26.

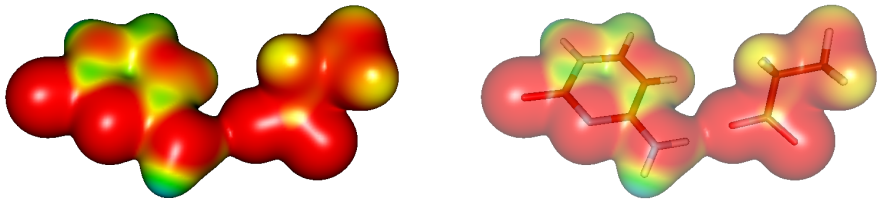


Figure A.36: Electrostatic potential of the sequence-specific cytosine – aspartate pair. Color coding as in Figure A.26.

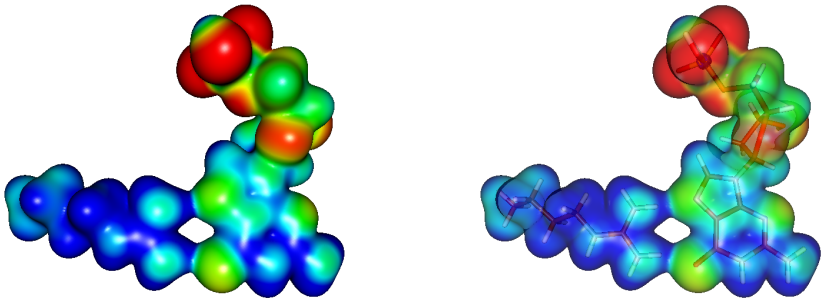


Figure A.37: Electrostatic potential of the sequence-specific dGMP – arginine pair. Color coding as in Figure A.26.

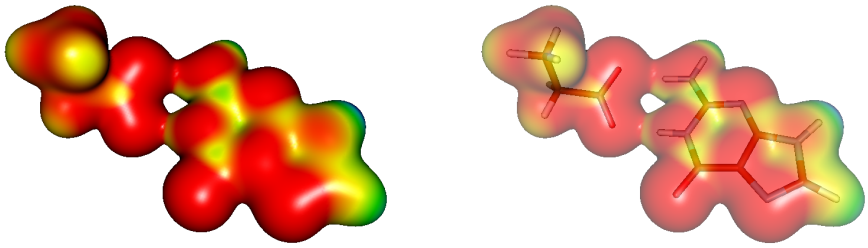


Figure A.38: Electrostatic potential of the sequence-specific guanine – aspartate pair. Color coding as in Figure A.26.

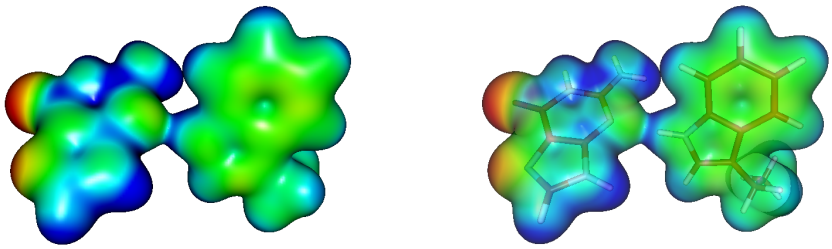


Figure A.39: Electrostatic potential of the sequence-specific guanine – tryptophan pair. Color coding as in Figure A.26.

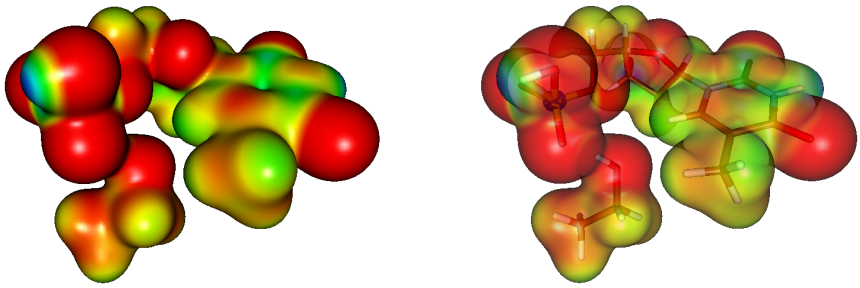


Figure A.40: Electrostatic potential of the sequence-specific TMP – serine pair. Color coding as in Figure A.26.

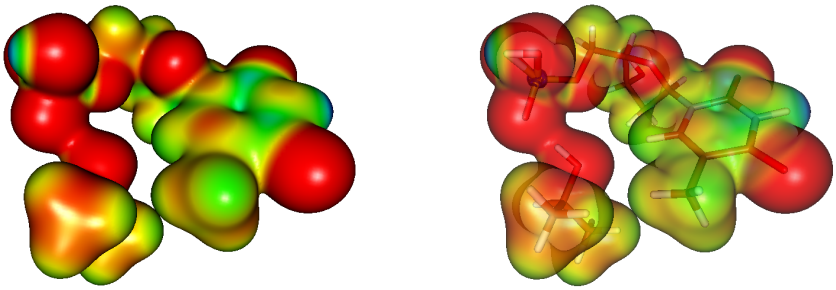


Figure A.41: Electrostatic potential of the sequence-specific TMP – threonine pair. Color coding as in Figure A.26.

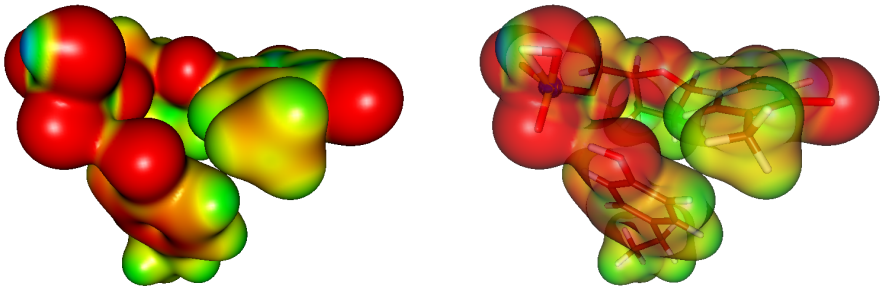


Figure A.42: Electrostatic potential of the sequence-specific TMP – tyrosine pair. Color coding as in Figure A.26.

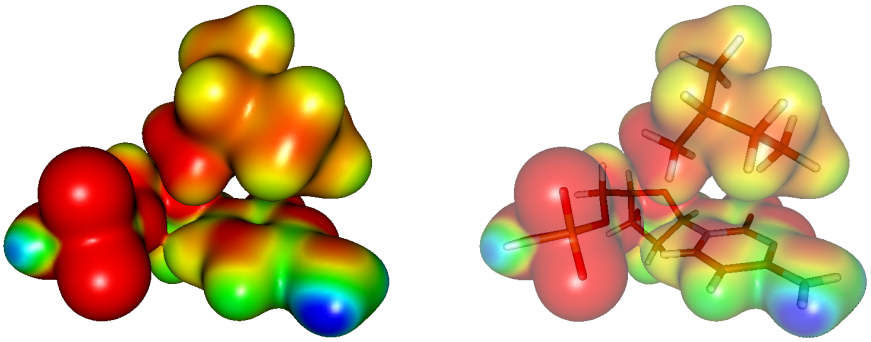


Figure A.43: Electrostatic potential of the sequence-specific dCMP – isoleucine pair. Color coding as in Figure A.26.

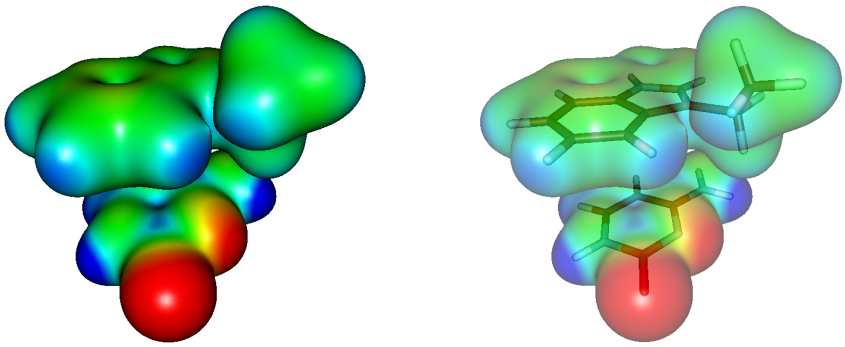


Figure A.44: Electrostatic potential of the sequence-specific cytosine – tryptophan pair. Color coding as in Figure A.26.